KTH Electrical Engineering

# Learning Stochastic Nonlinear Dynamical Systems Using Non-stationary Linear Predictors

MOHAMED RASHEED-HILMY ABDALMOATY

Licentiate Thesis
Stockholm, Sweden
2017

KTH Royal Institute of Technology
School of Electrical Engineering
Department of Automatic Control
SE-100 44 Stockholm
SWEDEN

TRITA-EE 2017:172
ISSN 1653-5146
ISBN 978-91-7729-624-9

Akademisk avhandling som med tillstånd av Kungliga Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie licenciatexamen i elektro- och systemteknik onsdag den 20 december 2017 klockan 10.00 i Hörsal Q2 (rumsnr B:218), Osquldas väg 10, Q-huset, våningsplan 2, KTH Campus, Stockholm.

Tryck: Universitetsservice US AB

# Abstract

Nonlinear stochastic parametric models are widely used in various fields whenever linear models are not adequate. However, the problem of parameter identification is very challenging for such models; the likelihood function – which is the central object in statistical inference – is, in general, analytically intractable. This renders favored point estimation methods, such as the maximum likelihood method and the prediction error methods, analytically intractable. In recent years, several methods have been developed to approximate the maximum likelihood estimator and the optimal mean-square error predictor using Monte Carlo approximation methods. However, the available algorithms can be computationally expensive; their application is so far limited to cases where fundamental difficulties, such as sample degeneracy and impoverishment problems, can be avoided.

The contributions of this thesis can be divided into two main parts. In the first part, several approximate solutions to the maximum likelihood problem are explored. Both analytical and numerical approaches, based on the expectation-maximization algorithm and the quasi-Newton algorithm, are considered. The performance of the developed algorithms is demonstrated on several numerical examples highlighting their advantages and disadvantages. These examples show that an analytic approximation of the likelihood function using Laplace's method may be acceptable; the accuracy depends not only on the true system, but also on the used input signal. While analytic approximations are difficult to analyze, asymptotic guarantees can be established for methods based on Monte Carlo approximations. Yet, Monte Carlo methods come with their own computational difficulties. Sampling in high-dimensional spaces requires an efficient proposal distribution. The main challenge is to reduce the number of Monte Carlo samples to a reasonable value while ensuring that the variance of the approximation is small enough.

In the second part, relatively simple prediction error method estimators are proposed. They are based on non-stationary one-step ahead predictors which are linear in the observed outputs, but may be nonlinear in the (assumed known) input. These predictors rely only on the first two moments of the model and the computation of the likelihood function is not required. Consequently, the resulting estimators are defined by analytically tractable objective functions in several relevant cases. The optimal linear one-step ahead predictor is derived and a corresponding prediction error estimator is defined. It is shown that, under mild assumptions, the classical asymptotic limit theorems are applicable and therefore the estimators are consistent and asymptotically normal. In cases where the first two moments are analytically intractable due to the complexity of the model, it is possible to resort to vanilla Monte Carlo approximations. Several numerical examples demonstrate a good performance of the suggested estimators in several cases that are usually considered challenging.

# Sammanfattning

Icke-linjära parametriska modeller används vid många tillämpningar för att modellera stokastiska system där linjära modeller inte anses vara adekvata. Dessa är dock inte helt triviala att tillämpa eftersom paramateridentifiering i dessa modellstrukturer är väldigt komplicerat; likelihood-funktionen – ett centralt objekt i statistisk inferens – är, generellt sett, analytiskt svårbehandlad. Detta medför att välanvända metoder, såsom maximum likelihood- och prediktionsfelmetoder, blir svårtillämpade. Diverse metoder har utvecklats under de senaste åren för att approximera maximum likelihood-skattaren och den optimala prediktorn (med avseende på medelkvadratsfelet) med hjälp av Monte Carlo metoder. Dock så kan dessa algorimer vara beräkningsmässigt krävande, och deras tillämpningsområden är idag begränsade till fall där fundamentala svårigheter, såsom partikeldegeneration, kan undvikas.

I den här avhandlingen behandlar vi det ovan beskrivna problemet via två tillvägagångssätt. I den första delen utforskas approximativa metoder för att lösa maximum likelihood-problemet. Både analytiska och numeriska metoder, baserade på expectation-maximization algoritmen och quasi-Newton metoden, behandlas. Vi demonstrerar metodernas effektivitet, samt tillkortakommanden, i numeriska simulationer. Exemplerna visar att analytiska approximationer av likelihood-funktionen (via Laplaces metod) kan vara acceptabla; nogrannheten beror inte enbart på det sanna systemet, utan även på insignalen som använts. Till skillnad från metoder baserade på Monte Carlo approximationer, för vilka asymtotiska garantier kan ges, är de analytiska approximationerna svåra att analysera. Monte Carlo metoder dras dock med vissa beräkningsmässiga svårheter; t.ex. är det svårt att dra sampel i högdimensionella rum eftersom det kräver en bra förslagsdistribution. Den huvudsakliga utmaningen som vi ställs inför är att reducera antalet Monte Carlo-sampel som krävs (till en resonlig nivå), medan variansen hos skattningen hålls låg.

I den andra delen av avhandlingen förslås relativt simpla skattare baserade på prediktionsfelsmetoden. Dessa använder enstegsprediktorn, som är linjär i den observerade utsignalen, men kan vara icke-linjär i insignalen (som antas vara känd). Dessa prediktorer *i)* använder enbart de första två momenten hos modellen, och *ii)* kräver inte att likelihood-funktionen beräknas. Detta medför att de föreslagna skattarna ger upphov till analytiskt hanterbara kostnadsfunktioner för ett flertal relevanta modellstrukturer. Den optimala linjära enstegsprediktorn härleds och den motsvarande prediktionsfelsskattaren defineras. Vi visar att klassiska resultat inom asymtotisk statistik är tillämpbara, under milda antaganden, vilket medför att vi kan visa att skattarna är konsistenta och asymptotiskt normalfördelade. I de fall där modellens första två moment är analytiskt komplicerade kan man falla tillbaka på vanliga Monte Carlo approximationer. Vi demonstrerar, via numeriska simulationer, att de föreslagna skattarna presterar bra för ett flertal exempel som oftast anses svårbehandlade.

إلي     أمي الغالية أ. ليلى فارس،
أبي الحبيب أ.د. رشيد حلمي.

# Acknowledgements

I would like to express my sincere gratitude to Håkan Hjalmarsson, my main supervisor, for his support and valuable guidance throughout the work of this thesis; I have learned and am still learning a lot, thank you Håkan!

I am also grateful to Cristian Rojas, my co-supervisor, for always being available for fruitful discussions, and want to thank him for his valuable time and advice.

My thanks also go to my former and current colleagues at the Automatic Control department for the stimulating work environment and friendly companionship; particularly to Robert Mattila for translating the abstract into Swedish.

Finally, I would like to thank my family for their love, moral support and encouragement; I am indebted to you!

*Mohamed Rasheed-Hilmy Abdalmoaty*
November 2017
Stockholm, Sweden

# Contents

# Notations

A bold font is used to denote random variables and a regular font is used to denote realizations thereof. The symbol ■ is used to terminate proofs.

### Number Sets

| | |
|---|---|
| $\mathbb{N}$, $\mathbb{N}_0$ | the set of natural numbers, $\mathbb{N} := \{1, 2, \dots\}$, $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. |
| $\mathbb{Z}$ | the set of integers, $\mathbb{Z} := \{0, \pm 1, \pm 2, \dots\}$. |
| $\mathbb{R}$ | the set of real numbers. |
| $\mathbb{R}_+$ | the set of nonnegative real numbers. |
| $\mathbb{R}^d$ | the Euclidean real space of dimension $d \in \mathbb{N}$. |

### Parameters and Constants

| | |
|---|---|
| $\theta$ | the parameter: a finite dimensional vector in $\mathbb{R}^d$ to be estimated. |
| $\theta^\circ$ | the (assumed) true parameter. |
| $\Theta$ | a compact subset of $\mathbb{R}^d$. |
| $\{g_k\}_{k=1}^{\infty}$ | impulse response sequence of a plant model. |
| $\{h_k\}_{k=0}^{\infty}$ | impulse response sequence of a noise model. |
| $d_w$, $d_u$, $d_y$ | the dimension of the process disturbance, the input signal and the output signal respectively; all belong to $\mathbb{N}$. |

### Signals and Stochastic Processes

| | |
|---|---|
| $\boldsymbol{\zeta} = \{\boldsymbol{\zeta}_t\}_{t \in \mathbb{Z}}$ | a generic vector-valued discrete-time stochastic process. |
| $\boldsymbol{x} = \{\boldsymbol{x}_t\}_{t \in \mathbb{Z}}$ | a latent/state process. |
| $u = \{u_t\}_{t \in \mathbb{Z}}$ | the input signal. |
| $\boldsymbol{y} = \{\boldsymbol{y}_t\}_{t \in \mathbb{Z}}$ | the output signal. |
| $\boldsymbol{w} = \{\boldsymbol{w}_t\}_{t \in \mathbb{Z}}$ | the process disturbance. |
| $\boldsymbol{v} = \{\boldsymbol{v}_t\}_{t \in \mathbb{Z}}$ | the measurement noise (when it is white). |
| $\boldsymbol{e} = \{\boldsymbol{e}_t\}_{t \in \mathbb{Z}}$ | the prediction error process. |
| $\boldsymbol{\varepsilon} = \{\boldsymbol{\varepsilon}_t\}_{t \in \mathbb{Z}}$ | the (linear) innovations process. |
| $\Phi_\zeta(\omega)$ | the power spectrum of a process $\boldsymbol{\zeta}$, in which $\omega \in \mathbb{R}$. |

### Vectors and Matrices

| | |
|---|---|
| $A(\theta)$ | state matrix of linear state space models parameterized by $\theta$. |
| $B(\theta)$ | input matrix of linear state space models parameterized by $\theta$. |
| $C(\theta)$ | output matrix of linear state space models parameterized by $\theta$. |
| $\boldsymbol{Z}$ | a column vector, $\boldsymbol{Z} := [\boldsymbol{\zeta}_1^\top, \ldots, \boldsymbol{\zeta}_N^\top]^\top$. |
| $\boldsymbol{X}$ | a column vector stacking state vectors, $\boldsymbol{X} := [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top]^\top$. |
| $U_t$ | a column vector stacking input vectors, $U := [u_1^\top, \ldots, u_t^\top]^\top$. |
| $\boldsymbol{Y}_t$ | a column vector stacking output vectors, $\boldsymbol{Y}_t := [\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_t^\top]^\top$. |
| $U, \boldsymbol{Y}$ | denote $U_N, \boldsymbol{Y}_N$ respectively. |
| $\boldsymbol{W}$ | a column vector stacking disturbance vectors, $\boldsymbol{W} := [\boldsymbol{w}_1^\top, \ldots, \boldsymbol{w}_N^\top]^\top$. |
| $\boldsymbol{V}$ | a column vector stacking measurement noise vectors, $\boldsymbol{V} := [\boldsymbol{v}_1^\top, \ldots, \boldsymbol{v}_N^\top]^\top$. |
| $\boldsymbol{E}$ | a column vector stacking prediction error vectors, $\boldsymbol{E} := [\boldsymbol{e}_1^\top, \ldots, \boldsymbol{e}_N^\top]^\top$. |
| $\boldsymbol{\mathcal{E}}$ | a column vector stacking innovation vectors, $\boldsymbol{\mathcal{E}} := [\varepsilon_1^\top, \ldots, \varepsilon_N^\top]^\top$. |
| $\hat{\boldsymbol{y}}_{t\|t-1}$ | one-step ahead predictor of $\boldsymbol{y}$. |
| $\widehat{\boldsymbol{Y}}$ | a column vector stacking one step ahead predictors of $y_t$, $\widehat{\boldsymbol{Y}} := [\hat{\boldsymbol{y}}_{1\|0}^\top, \ldots, \hat{\boldsymbol{y}}_{N\|N-1}^\top]^\top$. |
| $\hat{\theta}$ | estimate of $\theta$. |

### Functions and Operators

| | |
|---|---|
| $(\cdot)^{-1}$ | the inverse (of an operator or a square matrix). |
| $(\cdot)^\top$ | the transpose (of a vector or a matrix). |
| $F^\top(\theta), \Sigma^{-1}(\theta)$ | means $[F(\theta)]^\top, [\Sigma(\theta)]^{-1}$ respectively. |
| $\mathbb{1}_S(\cdot)$ | the indicator function of a set $S$. $\mathbb{1}_S(x) = 1$ if $x \in S$ and 0 otherwise. |
| $\\|\cdot\\|_2,\ \\|\cdot\\|_Q$ | the Euclidean norm and a weighted Euclidean norm. For any vector $x \in \mathbb{R}^n$, $n \in \mathbb{N}$ and any positive definite matrix $Q$, $\\|x\\|_Q := \sqrt{x^\top Q x}$. |
| $q$ | the shift operator on sequence spaces, $q x_t := x_{t+1}$ . |
| $q^{-1}$ | the backward shift operator on sequence spaces, $q^{-1} x_t := x_{t-1}$. |
| $G(q, \theta)$ | transfer operator plant model parameterized by $\theta$. |
| $H(q, \theta)$ | transfer operator noise model parameterized by $\theta$. |
| $f, g$ and $h$ | static (measurable) functions between Euclidean spaces. |
| $\psi$ | one-step ahead predictor function defining $\hat{\boldsymbol{y}}_{t\|t-1}$. |
| $\ell$ | scalar function used in the definition of prediction error methods, $(\varepsilon, t, \theta) \mapsto \ell(\varepsilon, t, \theta) \in \mathbb{R}_+$. |

## Probability Spaces

| | |
|---|---|
| $(\Omega, \mathcal{F}, P_\theta)$ | generic underlying measure space. The measure is parameterized by $\theta$. |
| $\delta_{\tilde{x}}(\mathrm{d}x)$ | Dirac measure supported at a point $\tilde{x}$. |
| $\mathbb{E}[\cdot;\theta]$ | mathematical expectation with respect to $P_\theta$. |
| $\mathbb{E}_{\boldsymbol{\zeta}}[\cdot;\theta]$ | mathematical expectation with respect to the distribution of a random vector $\boldsymbol{\zeta}$. The distribution is parameterized by $\theta$. |
| $p(\boldsymbol{\zeta};\theta)$ | probability density function of a real vector-valued random variable $\boldsymbol{\zeta}$ parameterized by $\theta$ with respect to the Lebesgue measure. |
| $p_{\boldsymbol{\zeta}}(\zeta;\theta)$ | the value $p(\boldsymbol{\zeta} = \zeta;\theta)$. |
| $p(\zeta;\theta)$ | the likelihood function of $\theta$ given a realization of $\boldsymbol{\zeta}$. |
| $p(\boldsymbol{\zeta}\vert\eta;\theta)$ | conditional probability density function, parameterized by $\theta$, of a real vector-valued random variable $\boldsymbol{\zeta}$ given a realization of another random variable $\boldsymbol{\eta}$; that is $p(\boldsymbol{\zeta}\vert\boldsymbol{\eta} = \eta;\theta)$. The reference measure is always the Lebesgue measure. |
| $\boldsymbol{\zeta}\vert\eta$ | a (conditional) random variable $\boldsymbol{\zeta}$ with a PDF $p(\boldsymbol{\zeta}\vert\eta;\theta)$. |
| $\mathbb{E}[\cdot\vert\eta;\theta]$ | conditional expectation given that $\boldsymbol{\eta} = \eta$. |
| $\mathbf{cov}(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2;\theta)$ | the covariance matrix of two random vectors $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2$; that is $\mathbb{E}\left[(\boldsymbol{\zeta}_1 - \mathbb{E}[\boldsymbol{\zeta}_1;\theta])(\boldsymbol{\zeta}_2 - \mathbb{E}[\boldsymbol{\zeta}_2;\theta])^\intercal;\theta\right]$. |
| $\mathbf{var}(\boldsymbol{\zeta};\theta)$ | the variance of a real-valued random variable $\boldsymbol{\zeta}$. |
| $\rightsquigarrow$ | means "converges in distribution to". |
| $\xrightarrow{\text{a.s.}}$ | means "converges almost surely to". |
| $\hat{\boldsymbol{\theta}}$ | estimator of $\theta$; that is $\hat{\boldsymbol{\theta}} \coloneqq \hat{\theta}(\boldsymbol{D}_N)$. |

## Function Spaces

| | |
|---|---|
| $\mathsf{L}_2(\Omega, \mathcal{F}, P_\theta)$ | the Hilbert space of real-valued random variables with finite second moments. Generally, the arguments will be dropped and only $\mathsf{L}_2$ will be used. |
| $\mathsf{L}_2^n(\Omega, \mathcal{F}, P_\theta)$ | the Hilbert space of random variables in $\mathbb{R}^n$ with entries in $\mathsf{L}_2$, and $n \in \mathbb{N}$. Generally, the arguments will be dropped and only $\mathsf{L}_2^n$ will be used. |
| $\varphi$ | a generic (measurable) test function. |
| $\langle\cdot,\cdot\rangle$ | the inner product of the Hilbert space $\mathsf{L}_2(\Omega, \mathcal{F}, P_\theta)$ , for any $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{L}_2$, the inner product $\langle\boldsymbol{x}, \boldsymbol{y}\rangle \coloneqq \mathbb{E}[\boldsymbol{x}\boldsymbol{y};\theta]$. |
| $\mathbf{sp}\{S\}$ | the linear span of the subset $S \subset \mathsf{L}_2$. |
| $\mathcal{P}_{\mathcal{S}}[\cdot]$ | the orthogonal projection operator of $\mathsf{L}_2$ or $\mathsf{L}_2^n$ onto a closed subspace $\mathcal{S}$. The space is understood from the context. |

### Data Sets

| | |
|---|---|
| $N$ | the cardinality of the data set, $N \in \mathbb{N}$. |
| $\boldsymbol{D}_t$ | a set of input and output pairs, $D_t := \{(\boldsymbol{y}_k, u_k) : 1 \leq k \leq t\}$, $t \leq N$. |

### Standard Distributions

| | |
|---|---|
| $\mathcal{N}(\mu(\theta), \Sigma(\theta))$ | the multivariate Gaussian distribution with a mean vector $\mu(\theta)$ and a covariance matrix $\Sigma(\theta)$; both parameterized by $\theta$. |
| $\mathcal{U}([a, b])$ | the uniform distribution over the closed interval $[a, b]$. |

### Other

| | |
|---|---|
| $0$ | the zero vector (the space is understood from context). |
| $\infty$ | infinity. |
| $:$ | means “such that”. |
| $\forall$ | means “for all”. |
| $\sim$ | means “is distributed according to”. It is used in conjunction with probability measures, distribution functions or PDFs. |
| $\propto$ | means “is proportional to”. |
| $:=$ | means “is defined as”. |
| $\approx$ | means “is approximately equal to”. |
| $\mathbf{1}$ | a column vector of ones with the appropriate dimension. |
| $I$ | the identity matrix with the appropriate dimension. |
| $t$ | discrete time index for signals and time-dependent functions. |
| $z$ | complex variable of the $z$-transform. |
| $\omega$ | angular frequency variable. |
| $\arg\min_{\theta \in \Theta} f(\theta)$ | the set of global minimizers of a real-valued function $f$ over a compact set $\Theta$. |
| $\nabla_\theta f(\theta)$ | gradient of a real-valued function $f$. |
| $\nabla_\theta^2 f(\theta)$ | Hessian of a real-valued function $f$. |
| $\mathcal{O}(N)$ | a function such that $\lvert\mathcal{O}(N)\rvert \leq CN$, with nonnegative $C < \infty$, $N \in \mathbb{N}$. |
| $\log(\cdot)$ | the natural logarithm function, $\log : \mathbb{R}_+ \setminus \{0\} \to \mathbb{R}$. |
| $\Sigma_1 \geq \Sigma_2,\ \Sigma_1 \succ \Sigma_2$ | for any two symmetric matrices $\Sigma_1$ and $\Sigma_2$ with equal dimensions, means that $\Sigma := \Sigma_1 - \Sigma_2$ is a positive semidefinite matrix or a positive definite matrix respectively. |

# Abbreviations

| | |
|---|---|
| i.i.d. | Independent and Identically Distributed. |
| EM | Expectation-Maximization. |
| EnKF | Ensemble Kalman Filter. |
| KF | Kalman Filter. |
| LTI | Linear Time-Invariant. |
| L-PEM | Prediction Error Method based on the optimal linear predictor. |
| L-SPEM | Simulated Prediction Error Method based on the optimal linear predictor. |
| MAP | Maximum A Posteriori. |
| MC | Monte Carlo. |
| MCEM | Monte Carlo Expectation-Maximization. |
| MCMC | Markov Chain Monte Carlo. |
| ML | Maximum Likelihood. |
| MLE | Maximum Likelihood Estimate/Estimator. |
| MSE | Mean-Square Error. |
| OE-PEM | OE-type Prediction Error Method. |
| OE-SPEM | OE-type Simulated Prediction Error Method. |
| PDF | Probability Density Function. |
| PE | Prediction Error. |
| PEM(s) | Prediction Error Method(s). |
| WL-PEM | Prediction Error Method based on the optimal linear predictor and a “Gaussian log-likelihood form” criterion. |
| WL-SPEM | Simulated Prediction Error Method based on the optimal linear predictor and a “Gaussian log-likelihood form” criterion. |

Chapter 1

# Introduction

In this chapter, we introduce the topic of the thesis, motivate the problem, and highlight our contributions. The last section gives an outline of the thesis content.

## 1.1 Learning Dynamical Models

System identification is a scientific method concerned with learning dynamical models based on observed data (see [65, 92, 111, 138, 142]). It can be described as the joint activity of dynamical systems modeling and parameter estimation. Like any other scientific method, system identification is used to acquire new knowledge or correct and improve existing knowledge based on measurable evidence. In system identification, the evidence is given in terms of a set of measured signals (variables) known as the "data set". The mathematical model used to describe the relation between the measured signals constitutes the hypothesis of the method.

In engineering sciences, system identification is used as a tool for the design or the operation of engineering systems. For example, most of the modern control techniques and signal processing and fault detection methods are based on mathematical models obtained using system identification techniques.

### 1.1.1 Systems

The term "system" refers to any spatially and temporally bounded physical or conceptual object within which several variables interact to produce an observable phenomena (see [69, 95, 152]). The observable variables are called the outputs (or the output signals) of the system. We will assume here that the outputs reflect the behavior of the system in response to some external stimuli. The external variables that can be altered by an extraneous observer are called the inputs (or the input signals) of the system. All other external variables that cannot be altered by the observer are called disturbances (or disturbance signals). In some but not all cases, the disturbances can be directly observed (measured).

It is assumed here that a system follows some sort of causality. The inputs and the disturbances are considered to be the causes, and the outputs are the observable effect. This definition of a system is quite general and can accommodate many observable phenomena. For example a system can be an economic system, a human cell, the solar system, an electric motor, or an aircraft. It is possible to define inputs, disturbances and outputs for each of these systems. For instance, the solar system is affected by the gravity of neighboring stars which cannot be altered by the observer; such gravitational effect is therefore a disturbance. On the other hand, the behavior of an aircraft is influenced by the engine thrust that can be taken as an input, but is also influenced by gust which is a disturbance. For more complex systems, the discrimination between inputs, disturbances, and outputs becomes less clear.

In many scientific fields, including engineering sciences, most of the systems are dynamical systems with some sort of memory. The outputs of a dynamical system at a certain time do not only depend on the inputs and disturbances at the same time, but also on their entire history.

### 1.1.2 Mathematical Models

A fundamental step of any system identification procedure is the specification of a mathematical model set. A mathematical model is an abstract representation of a system in terms of a mathematical relation between its inputs, outputs and disturbances. In practice, a mathematical model is seen as an approximation of the real-life system's behavior, and cannot provide an exact description; consequently, one system can have several models under several assumptions and/or intended use.

Dynamical systems are usually modeled by a set of (partial or ordinary) differential or difference equations. Models corresponding to differential equations are called continuous-time models, while those corresponding to difference equations are called discrete-time models. When the coefficients of these equations are independent of time, the models are called time-invariant models.

A generic causal discrete-time model can be defined by the equation

$$y_t = f_t(\{u_k\}_{k=-\infty}^{t}, \{\zeta_k\}_{k=-\infty}^{t}; \theta_t)$$

in which $t \in \mathbb{Z}$ is some integer representing time, $y_t \in \mathbb{R}^{d_y}$ is a real vector representing the value of the outputs at time $t$, the sequence $\{u_k\}_{k=-\infty}^{t}$ represents the input history up to time $t$, in which $u_t \in \mathbb{R}^{d_u}$ for all $t \in \mathbb{Z}$ is a real vector representing the value of the inputs at time $t$, and the sequence $\{\zeta_k\}_{k=-\infty}^{t}$ represents the history of the disturbances up to time $t$ where $\zeta_t$ for all $t \in \mathbb{Z}$ is a real vector representing the values of the disturbances. The symbol $f_t$ denotes a generic mathematical function modeling the cause/effect relationship between the inputs and disturbances on one hand, and the outputs on the other. The function is parameterized by a parameter $\theta_t$. Such a parameter is usually a finite-dimensional real vector, $\theta_t \in \mathbb{R}^d$ for all $t \in \mathbb{Z}$ and some $d \in \mathbb{N}$, in which case the model is said to be parametric. The subscript $t$ indicates that the function and the parameter may, in general, vary with time.

A model may be derived using physical laws and prior knowledge about the system. However, in cases where physical modeling is not possible due to the complexity of the system, standard classes of models can be used. An important subset of models is the set of parametric linear time-invariant models. These are models that assume a linear relationship between the inputs, the disturbances and the outputs such that the parameter is time-independent: $\theta_t = \theta$ for all $t \in \mathbb{Z}$ (see Section 2.1.3). Linear models are used extensively in practice, even when the underlying system exhibits a nonlinear behavior. This is mainly due to the fact that the estimation and feedback control theory is well developed and understood for linear models (see [6, 7, 53, 67, 68, 114, 137, 153]). However, when linear models are not accurate enough for the intended use, nonlinear models have to be considered (see [130] for a related discussion).

### 1.1.3 Estimation Methods

Once a model set has been determined, the next step of the system identification procedure is to choose a parameter estimation method. The choice is guided by the available assumptions on the data and the model class. The main goal of the estimation method is the evaluation of the unknown parameter vector $\theta$. An estimate is usually computed by solving either an optimization problem or an algebraic problem based on a set of recorded input and output signals over a finite time interval. Furthermore, the estimation method must provide some kind of accuracy measure for the computed estimate.

Since the values of the disturbances are usually not given, an uncertainty concept must be introduced. There are two main approaches for the characterization of uncertainty: the unknown but bounded approach, and the stochastic approach. In the unknown but bounded approach (see [100]), the uncertainty is characterized by defining a membership set for all the uncertain quantities. That is, for all $t \in \mathbb{Z}$, the values assumed by $\zeta_t$ belong to some known bounded set. Based on this constraint and the given data, a set of feasible parameters can be determined and a parameter can be selected by minimizing the worst-case error according to some performance measure. This approach is known by the name of “worst-case identification”. The stochastic approach, on the other hand, assumes a random nature of the uncertainty which is characterized by some probability distribution.

In this thesis, we will only consider the stochastic approach. The estimation step is then seen as an application of statistical inference methods. Under the assumption of a “frequentist” (see [82]) stochastic framework, the analysis of identification methods investigates what would happen if the experiment was to be repeated. The result of a “good” method is expected, for example, not to vary significantly. The analysis also examines what would happen to the result if very long (“infinite”) data records are available. It is important to understand that, even though only finite data records are available and even if the experiment is performed only once, the answers to such questions give confidence in the estimation method and are also used to compare and choose between different available estimation methods.

The most commonly used statistical estimation methods in system identification are the Maximum Likelihood (ML) method and the Prediction Error Methods (PEM) based on Prediction Error (PE) minimization (see [19, 52, 92, 111, 138]). Both are instances of a wider class of estimators known as the class of Extremum Estimators (see, for example, [1, Chapter 4]). Estimators in this class are defined by maximizing or minimizing a general objective function of the data and the parameter. The result is a point estimate, i.e., a single value for the parameter, together with an approximate confidence region.

In the case of the ML method, the objective function to be maximized is the likelihood function of the parameter. It is defined by the joint Probability Density Function (PDF) of the possible model outputs over some interval of time. The likelihood function is equal to the value of the PDF evaluated at the observed outputs and seen as a function of the parameter. Accordingly, to be able to compute the Maximum Likelihood Estimate (MLE) for a given model, it should be possible to compute the required joint PDF. Unfortunately, for general nonlinear models, this task is not trivial because the likelihood function has no analytic form.

On the other hand, the objective function to be minimized in the case of PEMs is given by the sum of the errors made by the model when used to predict the observed outputs. This requires the definition of: (i) an output predictor function based on the assumed statistical model, and (ii) a distance measure in the output signal space to evaluate the errors. Different choices for the predictor functions and the distance measure lead to different instances of the family of PEMs. An optimal PEM instance, in the sense of minimizing the expected squared prediction errors, can be defined. However, it relies indirectly on the joint PDF of the possible model outputs. Consequently, for general nonlinear models, the objective function of the optimal PEM instance has no analytic form.

### 1.1.4 Properties of Estimators

The ML method and the PEMs are favored due to their statistical properties. To be able to discuss and compare statistical properties of estimators, we usually need the assumption of a true system. Namely, we assume that there exists a "true" parameter, denoted by $\theta^\circ$, such that the recorded observation is a realization of the outputs of the model with $\theta = \theta^\circ$. The model evaluated at $\theta^\circ$ is said to be the "true model/system". Although such an assumption is never true in practice, it is convenient for the theoretical analysis of the estimation methods.

Usually, the considered properties of estimators are asymptotic in nature. Perhaps the weakest property that should be required for any estimation method is "consistency". Consistency is a central idea in statistics; it means that as the data size increases, the resulting estimates become closer to the true parameter. Because an estimator is a random variable (a function of random variables), such a convergence is taken in a probabilistic sense (see [23, Chapter 4]). For example, convergence can be defined as follows: for an arbitrary probability $P \in [0, 1]$ close to 1 and any topological ball centered at $\theta^\circ$ with an arbitrary small radius, we only need data records long enough so that the estimate of $\theta$ is inside the given ball with probability $P$. We then say that the estimator converges in probability to the true parameter as the data size grows towards infinity. An estimator with such a property is called a (weakly) consistent estimator of $\theta$.

Efficiency is another property used to compare consistent estimators. Given two consistent estimators of $\theta$, it is natural to use the one that gives estimates closer to the true parameter. This can be evaluated by comparing the (asymptotic) distributions of a normalized version of the errors $(\hat{\boldsymbol{\theta}} - \theta^{\circ})$. A consistent estimator is called asymptotically efficient if its normalized error has an asymptotic covariance no larger than any other consistent estimator. It is usually the case that $\lim_{N\to\infty} \sqrt{N}(\hat{\boldsymbol{\theta}} - \theta^{\circ})$, in which $N$ is the data size and $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\theta$, has a multivariate Gaussian distribution with zero mean and some covariance matrix. It is then said that the estimator is asymptotically normal. Given several asymptotically normal estimators of the same parameter, the estimator with the smallest asymptotic covariance matrix should be preferred.

Under fairly general conditions, it can be shown that both the ML method and the PEMs lead to consistent and asymptotically normal estimators (see [19, 60, 92] for example). Furthermore, the MLE is asymptotically efficient under week assumptions. The PEM with the optimal predictor and a specific choice for the distance measure on the outputs can be shown to coincide with the MLE.

## 1.2 Motivation and Overview of Available Methods

In this thesis, we are concerned with the parameter estimation problem of fairly general stochastic nonlinear dynamical models. We are specifically interested in cases where an unobserved disturbance or latent process is affecting the outputs through a non-invertible nonlinear transformation. This is illustrated in the block diagram in Figure 1.1. We will make the standing assumption that a parametric model set is given and we will not be concerned with the important step of model structure selection. Our objective is to apply either the ML method or a consistent instance of the PEMs.

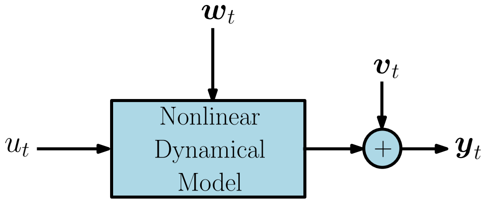


**Figure 1.1:** A stochastic nonlinear model. The input $u_t$ is known and the output $\boldsymbol{y}_t$ is a stochastic process. The disturbance $\boldsymbol{w}_t$ is an unobserved stochastic process affecting the output through the nonlinear dynamics, and $\boldsymbol{v}_t$ is an additive measurement noise. Both $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are usually assumed to be mutually independent and to have finite-dimensional distributions of known analytic forms.

To motivate the problem, we consider below two cases. In the first, the stochastic part of the outputs comes from an additive measurement noise; in this case, the model is invertible in the sense that, for any given $\theta$, the measurement noise can be reconstructed from the knowledge of the inputs and the outputs. In the second, a nonlinear contribution from an unobserved disturbance is also present; in this case the inputs and outputs cannot be used to reconstruct the unobserved disturbance, and the model is said to be non-invertible.

For each case, we provide a very brief overview of the available identification methods. There is a significant literature on this topic, and it is not possible to cover all the relevant work in a brief overview. However, we refer the interested reader to the surveys [12, 57, 126, 128, 134], the books [13, 48, 101, 104, 111], and the exhaustive references lists therein. The Ph.D. theses [36, 87, 127, 132] cover the most recent methods dealing with identification for nonlinear systems.

### 1.2.1 Case 1: Invertible Models

Consider the static model

$$\boldsymbol{y}_t = g(u_t;\theta) + \boldsymbol{v}_t, \qquad t = 1, 2, 3, \dots \tag{1.1}$$

with scalar inputs $u_t$ and outputs $\boldsymbol{y}_t$, in which $g(\cdot;\theta) : \mathbb{R} \to \mathbb{R}$ is a nonlinear function parameterized by $\theta$, and for any time $t \in \mathbb{N}$ the random variable $\boldsymbol{v}_t$ has a known PDF, that is $\boldsymbol{v}_t \sim p(\boldsymbol{v}_t)$. Observe that a bold font is used to denote random variables and a regular font is used to denote realizations thereof. We assume that $\boldsymbol{v}_t$ and $\boldsymbol{v}_s$ are independent whenever $t \neq s$. Since the input sequence $\{u_t\}$ is known, it is clear that the outputs are independent over $t$.

Let us define the vector $\boldsymbol{Y}_t := [\boldsymbol{y}_1 \ \dots \ \boldsymbol{y}_t]^\top$, and let $\boldsymbol{Y} := \boldsymbol{Y}_N$. We can easily construct the joint PDF of the model outputs,

$$p(\boldsymbol{Y};\theta) = \prod_{t=1}^{N} p(\boldsymbol{y}_t;\theta) = \prod_{t=1}^{N} p_{\boldsymbol{v}_t}(\boldsymbol{y}_t - g(u_t;\theta)), \tag{1.2}$$

and therefore we have no trouble formulating the ML optimization problem (see Definition 2.4.2):

$$\hat{\boldsymbol{\theta}}_{\text{ML}} := \arg\max_{\theta} \prod_{t=1}^{N} p_{\boldsymbol{v}_t}(\boldsymbol{y}_t - g(u_t;\theta)).$$

Because the outputs are independent over time, it is also easy to show, see [92, Chapter 3], that the optimal (mean-square error) one-step ahead predictor is given by

$$\hat{y}_{t|t-1}(\theta) := g(u_t;\theta),$$

and we can simply define a suboptimal but consistent (see Definitions 2.3.2 and 2.3.3) PEM estimator (see Definition 2.4.4) as the minimizer of an unweighted nonlinear least-squares problem:

$$\hat{\boldsymbol{\theta}}_{\text{PEM}} := \arg\min_{\theta} \sum_{t=1}^{N} (\boldsymbol{y}_t - \hat{y}_{t|t-1}(\theta))^2 = \arg\min_{\theta} \sum_{t=1}^{N} (\boldsymbol{y}_t - g(u_t;\theta))^2.$$

Under some weak conditions on the model, the parameterization and the input signal (see [92, Chapter 8]), it is known that both estimators are consistent and asymptotically normal (see Definition 2.3.4).

We get a direct extension to the dynamic case if we allow the function $g$ to depend on previous inputs and outputs. Most of the classical research found in the literature on nonlinear system identification is dedicated to this case, and generally focuses on two main issues: (i) the problem of model selection and parameterization, and (ii) the optimization methods used to fit the parameters to the data. Most of the work is done using a slightly more general model compared to (1.1) and is given by the equations

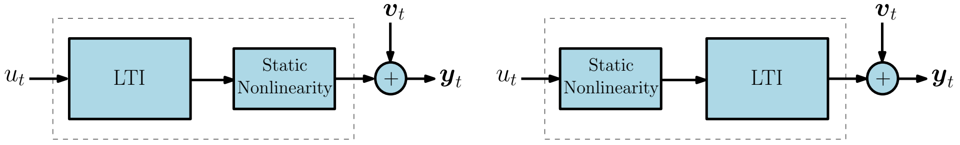
$$\boldsymbol{y}_t = \psi(\boldsymbol{\varphi}(t;\theta);\theta) + \boldsymbol{v}_t, \quad t = 1, 2, 3, \dots \tag{1.3}$$

The nonlinear mapping $\psi$ is known as the predictor function, and $\boldsymbol{\varphi}(t;\theta)$ is a parameterized regression vector that is a function of past inputs, past outputs, and past prediction errors $\boldsymbol{e}_t(\theta) := \boldsymbol{y}_t - \psi(\boldsymbol{\varphi}(t;\theta);\theta)$. Observe that the model in (1.3) assumes that $\boldsymbol{w}_t$ in Figure 1.1 is identically 0 for all $t$ but allows for a recurrent structure; i.e., the current output may depend on previous inputs and outputs (see [134]). Several possibilities of parameterizing the predictor function and selecting the regressor variables can be found in the above cited books and surveys. For instance, they include Volterra kernel representations, Nonlinear AutoRegressive eXogenous (NARX) models, Nonlinear AutoRegressive Moving Average eXogenous (NARMAX) models, nonlinear state-space models in predictor form, and block-oriented models with only additive measurement noise and no latent disturbances.

The Volterra representation is an example of a nonparametric structure and is described in detail in the book [124]. It can be seen as a generalization of the impulse response of linear models to the nonlinear case. Its identification requires the estimation of the kernels which, according to their orders, might contain thousands of parameters. The recent research effort in [14] applied regularization techniques to Volterra kernels estimation in the hope of improving the accuracy of the estimates for reasonable data sizes.

The NARX and NARMAX models are generalizations of the linear AutoRegressive eXogenous (ARX) and AutoRegressive Moving Average eXogenous (ARMAX) models defined in [92, Chapter 4]. They are flexible nonlinear model structures that provide input-output representations of a wide range of nonlinear systems including models with nonlinear feedback; they are studied in detail in the book [13]. One of the main advantages of such structures is their parsimony; the dimension of the parameterization vector of a NARX or a NARMAX model is small compared to other available nonlinear models.

Block-oriented models (see [48]) are models consisting of two main block types: static nonlinearity blocks and Linear Time-Invariant (LTI) dynamical blocks. The LTI blocks can be represented by parametric models such as rational transfer operators, linear state-space models, basis function expansions or by nonparametric models in either time or frequency domain. Similarly, the static nonlinearity can be represented by a nonparametric kernel model, or by some linear-in-parameters basis function model. A block-oriented model structure is developed by connecting several of these two building blocks together in series or in parallel, or both. Feedback connections are also possible. Although not as general as the Volterra representation or the NARMAX models, the block-oriented structures may be used to model many real-life nonlinear systems (see [48]). The simplest block-oriented structures are series connections involving only one block of each type, see Figure 1.2. A Wiener model is constructed by a single LTI block followed by a static nonlinearity block at the output; a Hammerstein model is constructed by a static nonlinearity block at the input followed by an LTI block. The model complexity can be increased by connecting these simple models in series or in parallel as shown in [48].



**(a)** Wiener model: an LTI model followed by a static nonlinearity at the output

**(b)** Hammerstein model: a static nonlinearity at the input followed by an LTI model.

**Figure 1.2:** A Wiener and a Hammerstein model.

One main advantage of block-oriented structures is the possibility of separating the estimation of the linear and nonlinear parts of the model. Under some assumptions on the input signal and the noise, it is possible to construct best linear approximations (BLA) of the model, see [36, 93, 129, 130], which can be shown to be related to the LTI blocks, see [128, 132]. The main tool there is Bussgang's theorem given in [18]; it states that the cross-correlation functions of a Gaussian signal before and after passing through a static nonlinear function are equal up to a constant. However, the constant may well be 0 (see Example 4.2.3 on page 102).

The choice among these different representations is usually guided by prior knowledge about the underlying system, the intended use of the model, and the available computational resources, among others. The selection of the model structure is fundamental and is recognized to be the most difficult step in the system identification procedure. Several specialized methods of structure selection for nonlinear models have been developed in both time and frequency domain, see [13] and [128].

The parameter estimation step for the model in (1.3) remains relatively simple. The ML problem can be easily formulated assuming a known distribution for $\boldsymbol{v}_t$ and known initial conditions. Observe that the regression vectors of the NARMAX mode include delayed measurement noise; the NARMAX model is defined by the relations

$$\boldsymbol{y}_t = \psi(\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-n_y}, u_{t-1}, \ldots, u_{t-n_u}, \boldsymbol{v}_{t-1}, \ldots, \boldsymbol{v}_{t-n_v}; \theta) + \boldsymbol{v}_t, \quad t \in \mathbb{Z},$$

for some $n_y, n_u, n_v \in \mathbb{N}$. The values $\boldsymbol{v}_{t-1}, \ldots, \boldsymbol{v}_{t-n_v}$ can be evaluated using the model, the initial conditions and the previous inputs and outputs (it is an invertible model). This allows us to compute the likelihood function in closed-form. Similarly, it is still possible to formulate a PEM problem in closed-form. A PEM estimator can be defined, for example, as the global minimizer

$$\hat{\boldsymbol{\theta}} := \arg\min_{\theta\in\Theta} \sum_{t=1}^{N} \|\boldsymbol{y}_t - \psi(\boldsymbol{\varphi}(t;\theta);\theta)\|^2.$$

Depending on the parameterization of the predictor function and the regression vector, the solution is usually not available in closed-form. The resulting optimization problem is in general non-convex. Nonlinear optimization algorithms like the damped Gauss-Newton algorithm or the Levenberg-Marquardt algorithm (see [42, 104, 108]) are therefore required. These algorithms are numerical iterative methods that can only find a local minimum. They require a good initial value to guarantee that the solution is in the vicinity of a global minimizer. In the case of block-oriented models, linear approximations can be used to initialize the optimization algorithms. There have been some research efforts in this direction, see for example [97, 110, 132, 133, 135, 136]. However, it should be noted that the problem of finding a linear approximation, depending on the chosen model structure for the LTI blocks, might still be a difficult non-convex problem.

To avoid possible local minima for complicated parameterizations, it is also possible to use random global search strategies such as simulated annealing and genetic algorithms (see [104, Chapter 5] or [117, Chapter 5]). However, due to their random nature, the exploration of the entire optimization domain could be computationally expensive and time-consuming. On the other hand, several approaches constrain the possible parameterization of the model in such a way that the resulting optimization problem has a closed-form solution (for example considering only linear-in-parameters models (see [134, Section 8] or [13, Chapter 3]).

The model in (1.3) ignores any possible stochastic process (different from $\boldsymbol{v}_t$) that passes through the nonlinear dynamics. This is an idealization of the real situation where there might exist a disturbance entering the system before some nonlinear subsystem. It has been shown in [58] that if the process disturbance is ignored, the resulting estimators will not be consistent. Therefore, it is important to develop identification methods that take into account the presence of such disturbances, which is the aim of this thesis.

### 1.2.2 Case 2: Non-invertible Models

Assume that the output does not only depend on the input and the variables $\boldsymbol{v}_t$, but also on some (latent) unobserved process $\boldsymbol{w}_t$ through a non-invertible (in $\boldsymbol{w}_t$) nonlinear function. In this case, the output becomes

$$\boldsymbol{y}_t = g(\boldsymbol{w}_t, u_t; \theta) + \boldsymbol{v}_t, \qquad t = 1, 2, 3, \dots \tag{1.4}$$

in which $g(\cdot,\cdot;\theta) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is a non-invertible (w.r.t. $\boldsymbol{w}_t$) nonlinear function parameterized by $\theta$.

Define the vector $\boldsymbol{W}_t := [\boldsymbol{w}_1 \;\ldots\; \boldsymbol{w}_t]^{\top}$, let $\boldsymbol{W} := \boldsymbol{W}_N$, and assume that $\boldsymbol{W} \sim p(\boldsymbol{W};\theta)$ which has a known functional form. Let us first consider the ML problem. Note that the outputs are independent only if the latent process $\boldsymbol{w}$ is independent over. We also observe that because $\boldsymbol{W}$ is not observed, the PDF of $\boldsymbol{Y}$ has to be calculated by marginalizing the joint distribution $p(\boldsymbol{Y},\boldsymbol{W};\theta)$ with respect to $\boldsymbol{W}$; that is

$$p(\boldsymbol{Y};\theta) = \int_{\mathbb{R}^{d_w N}} p(\boldsymbol{Y}, W;\theta)\,\mathrm{d}W. \tag{1.5}$$

Unfortunately, even though the integrand in (1.5) has a known form, the integral is a multidimensional integral over $\mathbb{R}^{d_w N}$ which, in general, has no closed-form expression. The symbol $d_w$ denotes the dimension of the latent process $\boldsymbol{w}$, which is assumed to be a scalar process in the current example (i.e., $d_w = 1$). For commonly encountered models and applications, the value of $d_w N$ is $\mathcal{O}(10^3)$ or $\mathcal{O}(10^4)$ and the evaluation of the integral in (1.5) is very challenging due to the nature of the integrand. To be able to find an approximate solution to the ML problem, one has to come up with computational methods that can approximate the maximizer of the intractable function $p(Y;\cdot)$ over a given set $\Theta$.

Currently, there is an ongoing research effort in this direction within the system identification community; see for example [86, 106, 125, 146, 147] for ML estimation based on sequential Monte Carlo (SMC) methods. The survey [126] summarizes the available state-of-the-art algorithms and distinguishes between two main approaches: (i) the marginalization approach, and (ii) the data augmentation approach.

In the marginalization approach, SMC filters and smoothers (also known as particle filters) are used to “marginalize out” the latent process to approximate the logarithm of the likelihood function (the log-likelihood) and its gradient at a given value of $\theta$. The resulting approximations are then used within an iterative numerical optimization algorithm to find an approximate solution to the ML problem. In the data augmentation approach, the SMC filters and smoothers are used in conjunction with the Expectation-Maximization (EM) algorithm (see [29] or Section 3.2.1) to approximate the MLE as suggested in [144]. The resulting Monte Carlo EM (MCEM) algorithm converges only if the number of the Monte Carlo samples (particles) grows with the algorithm iterations, see for example [20, 43]. Furthermore, a new set of particles has to be generated at each iteration of the algorithm. In order to make a more efficient use of the particles, [28] suggested the use of a stochastic approximation version of the EM algorithm (SAEM) which replaces the expectation step of the EM algorithm by one iteration of a stochastic approximation procedure. In [86], a particle Gibbs sampler (conditional particle filter with ancestor sampling, see [85]) is used within the SAEM algorithm to generate the required particles.

We note here that the available convergence proofs of the MCEM and the SAEM algorithms (see [28, 43]) are, so far, given for cases where $p(\boldsymbol{Y},\boldsymbol{W};\theta)$ belongs to the exponential family (see [123, Section 2.2] for a definition of the exponential family). This constrains the distributions of $\boldsymbol{w}$ and $\boldsymbol{v}$ as well as the parameterization of the model.

Estimation methods based on SMC approximations can be computationally expensive. For example, the convergence of the MCEM and the SAEM algorithms can be very slow when the variance of the latent process is small. Furthermore, cases with small or no measurement noise are considered challenging for SMC methods. Moreover, due to the sample degeneracy and impoverishment problems – a pair of fundamental difficulties of SMC methods (see [32, 33]) – the SMC methods are so far, to the best of the author's knowledge, applicable to models with relatively small $d_w$.

In this thesis, the ultimate goal is the construction of consistent estimators of $\theta$; we will be looking at alternative approximation methods that can be used to approximately solve the MLE problem without relying on an exact filtering or smoothing distribution. Instead, we only use the parameterized model which defines the joint PDF $p(\boldsymbol{Y}, \boldsymbol{W}; \theta)$. It is of interest to know if this point of view could lead to some computational advantage. We shall deal with this problem in Chapter 3.

The computation of the PEM estimator, for cases with non-invertible model, is not easier than the computation of the MLE. Observe that any estimator ignoring $\boldsymbol{w}$, as shown in [58], is not guaranteed to be consistent. So far, the available PEMs have relied on approximations of the optimal predictor. The optimal (mean-square error) one-step ahead predictor depends on the history of the observed outputs and is given by the conditional mean (see Chapter 2, page 34),

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) := \mathbb{E}[\boldsymbol{y}_t | \boldsymbol{Y}_{t-1}; \theta] = \int_{\mathbb{R}^{d_y}} y_t \, p(y_t | \boldsymbol{Y}_{t-1}; \theta) \, \mathrm{d}y_t, \quad t = 1, 2, 3, \ldots \tag{1.6}$$

in which $p(\boldsymbol{y}_t | \boldsymbol{Y}_{t-1}; \theta)$ is known as the predictive PDF and $\boldsymbol{Y}_0$ is defined as the empty set; hence $\hat{\boldsymbol{y}}_{1|0}(\theta)$ is constant and is equal to the expectation of $\boldsymbol{y}_1$. Unlike the MLE, the integrals appearing in the PEM objective function are of the same dimension as that of the output signal. This means that the domain of the integrand is independent of $N$. For single-input single-output models (like the model in (1.4)), these are one-dimensional integrals; however, the integrand has an unknown form. Unfortunately, to be able to calculate the predictive PDF, we need to compute multidimensional integrals. Observe that by Bayes' theorem (see [64, page 39]),

$$p(y_t | Y_{t-1}; \theta) = \frac{p(Y_t; \theta)}{\int_{\mathbb{R}^{d_y}} p(Y_t; \theta) \, \mathrm{d}y_t} = \frac{\int_{\mathbb{R}^{d_w t}} p(Y_t, W_t; \theta) \, \mathrm{d}W_t}{\int_{\mathbb{R}^{d_y}} \int_{\mathbb{R}^{d_w t}} p(Y_t, W_t; \theta) \, \mathrm{d}W_t \, \mathrm{d}y_t}. \tag{1.7}$$

It appears that the predictive PDF of the output depends on $p(Y_t; \theta)$ which does not have a closed-form expression. To be able to formulate and solve a PEM problem, one has to come up with either consistent instances of the PEMs without relying on the optimal predictor, or computational methods that can approximate the conditional mean (1.6). In the latter case, it seems that the PEM does not have any computational advantage over the efficient ML method. Both the MLE and the conditional mean of the outputs require the solution of similar marginalization integrals. For this reason, most of the recent research efforts found in the system identification literature, so far, target the maximum likelihood estimator.

One of the main contributions of this thesis is the introduction of one-step ahead predictors constructed using the postulated statistical model without any reference to the data or the the likelihood function (see Chapter 4). These predictors are relatively easy to compute and the can be used to construct consistent instances of the PEMs. The predictors do not necessarily coincide with the optimal predictor, which means that certain asymptotic properties (such as statistical efficiency) cannot be guaranteed. However, the introduced predictors are algorithmically and conceptually simpler than predictors based on either SMC smoothing algorithms or Markov Chain Monte Carlo (MCMC) algorithms.

### The main challenge

The two cases discussed above clarify the source of difficulty of the estimation problems considered in this thesis. The difference between the "tractable" model in (1.1) and the "intractable" model in (1.4) is the non-invertible nonlinear transformation of unobserved random variables in the latter. This makes the likelihood function and the optimal one-step ahead predictor of the output analytically intractable. Consequently, the objective functions of the optimization problems of both the MLE and the PEM estimator are not available in closed-form.

## 1.3 Thesis Outline and Contributions

The content of this thesis concerns the estimation problem of parametric nonlinear stochastic dynamical models. Firstly, several possible approximations of the MLE are explored. Secondly, computationally attractive PEM estimators based on non-stationary linear one-step ahead predictors are introduced. Below, an outline of each chapter is given where we also indicate the main contributions.

### Chapter 2: Background and Problem Formulation

Chapter 2 provides the necessary background for the thesis. Here, several important remarks and observations are made. After introducing a stochastic framework, a classical result on the structure of general second-order stochastic processes – due to Harold Cramér in [27] – is introduced. It generalizes Wold's decomposition (see [148]) of stationary processes by giving an exact description of the second-order properties of a non-stationary process in terms of a causal "linear" time-varying filtering of the innovation process. This interesting result is used in Chapter 4 as the basis for optimal linear prediction of non-stationary processes with nonlinear underlying models. The chapter also introduces two frequentist estimation methods: the ML method and the PEMs based on PE minimization. The kinship between the two methods is highlighted using linear dynamical models. Finally, the main problem of the thesis is formulated.

### Chapter 3: Approximate Solutions to the ML problem

In this chapter, we investigate several approaches to approximate solutions of the ML estimation problem. The focus is on the EM algorithm and the quasi-Newton algorithm. For both algorithms, analytical as well as numerical approximations are explored. The analytical approximations are based on a Gaussian approximation of the posterior of the unobserved (latent) disturbance (using Laplace's method). The numerical approximations, in cases where the output process is independent, may be obtained by using deterministic integration; however, in general, Monte Carlo approximations based on importance sampling are considered. The performance of the approximate algorithms is evaluated on several (relatively simple) numerical examples where the advantages and disadvantages of each method are highlighted.

The material in this chapter has not appeared in any publications, except the last part of Section 3.4.2 concerning Monte Carlo approximations of the quasi-Newton algorithm based on Laplace importance sampling (Algorithm 3.4.2); this algorithm has been published in

> Mohamed Abdalmoaty and Håkan Hjalmarsson. A Simulated Maximum Likelihood Method for Estimation of Stochastic Wiener Systems. In *the 55th IEEE Conference on Decision and Control (CDC)*, pp. 3060-3065, Las Vegas, USA, 2016

### Chapter 4: Linear Prediction Error Methods

In this chapter, we propose a relatively cost-efficient PEM based on suboptimal predictors. The used predictors are defined using only the first two moments of the postulated model; they are linear in the observed outputs, but are allowed to depend nonlinearly on the (assumed known) inputs. The optimal linear predictors, in the sense of minimum mean-square error, are derived and used to construct a PEM estimator. It is shown that for several relevant models with intractable likelihood functions (such as stochastic Wiener-Hammerstein models with polynomial nonlinearity), the suggested PEM estimators are defined by closed-form (exact) expressions. Under some mild assumptions, the resulting estimators are shown to be consistent and asymptotically normal. The chapter also discusses the relation between the proposed PEMs and a suggested approximation in Chapter 3. We give the PEM a maximum likelihood interpretation that allows for the use of the EM algorithm. The performance of the method is illustrated by numerical simulations using several challenging models. Finally, a comparison and a connection are made to a PEM based on the Ensemble Kalman filter – a Monte Carlo filter used to define a (nonlinear) suboptimal predictor.

The ideas developed in this chapter have originated in

> Mohamed Abdalmoaty and Håkan Hjalmarsson. Simulated Pseudo Maximum Likelihood Identification of Nonlinear Models. In *IFAC-PapersOnLine*, Volume 50, Issue 1, pp. 14058–14063, 2017.

### Chapter 5: Conclusions and Future Research Directions

In this last chapter, we summarize the conclusions of the thesis and give some pointers for future research.

### Appendices

The thesis contains three appendices where relevant definitions and results are summarized. Appendix A introduces the idea of Monte Carlo estimation. Random sampling, common random numbers, and importance sampling are defined and discussed. In Appendix B, Hilbert spaces of random variables are defined and the optimal linear mean-square error predictors are derived based on the projection theorem. Lastly, Appendix C gathers some relevant properties of Gaussian random vectors and multivariate Gaussian distributions.

Chapter 2

# Background and Problem Formulation

This chapter introduces the necessary background, formulates the main problem, and makes several remarks. We start by introducing a stochastic framework for the signals. We then describe the models that we are concerned with. Finally, we discuss statistical estimation methods and their properties.

## 2.1 Mathematical Models

The mathematical models considered in this thesis belong to the set of stochastic models; all the signals are modeled using stochastic processes. A stochastic process $\boldsymbol{y} = \{\boldsymbol{y}_t : t \in T\}$ is a family of random variables indexed by a given index set $T$ and defined over a common underlying probability space $(\Omega, \mathcal{F}, P_\theta)$. In this thesis, the probability measure is parameterized by a finite-dimensional real vector $\theta \in \Theta \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$. The index $t$ always refers to time and the index set $T$ is taken as the set of integers $\mathbb{Z}$ giving rise to discrete-time stochastic processes (a classical reference on the theory of stochastic processes is [31]). For any finite subset $\{t_1, \dots, t_N\} \subset \mathbb{Z}$, the joint distribution of the random variables $\{\boldsymbol{y}_{t_1}, \dots, \boldsymbol{y}_{t_N}\}$ is known as a finite-dimensional distribution of the process. For every $t$, we always assume the existence of a joint probability density function $p(\boldsymbol{Y}_t; \theta)$ of the vector $\boldsymbol{Y}_t := [\boldsymbol{y}_1^\top, \dots, \boldsymbol{y}_t^\top]^\top$. The probability measure $P_\theta$ can be characterized by specifying the finite-dimensional distribution for all finite subsets of $\mathbb{Z}$. The mathematical models are then deterministic objects that define the probability measure $P_\theta$ on the space of observed signals. Because all practical systems are causal systems, for which the current output does not depend on the future inputs or future disturbances, we limit ourselves to causal models. We start by defining special classes of signals.

### 2.1.1 Signals

The outputs, inputs, and disturbances are modeled using stochastic processes. The mathematical models to be developed are deterministic functions that define a mapping between these processes. We will only consider $d_\zeta$-dimensional real-valued second-order discrete-time stochastic processes $\boldsymbol{\zeta}$ with some finite $d_\zeta \in \mathbb{N}$.

**Definition 2.1.1** (Second-order discrete-time stochastic process). *A stochastic process $\boldsymbol{y} = \{\boldsymbol{y}_t : t \in \mathbb{Z}\}$ is said to be a second-order stochastic process if it holds that*

$$\begin{aligned} \mathbb{E}[\boldsymbol{y}_t] &= \mu_t, & \|\mu_t\| &\leq C < \infty, & \forall t &\in \mathbb{Z}, \\ \mathbb{E}[\boldsymbol{y}_t \boldsymbol{y}_s^\top] &= R_y(t,s), & \|R_y(t,s)\| &\leq C < \infty, & \forall t, s &\in \mathbb{Z} \end{aligned}$$

*where $C$ is a generic constant (that does not necessarily assume the same value when bounding different quantities).*

One of the simplest and most used classes of second-order stochastic processes is the class of stationary processes.

**Definition 2.1.2** (Stationary discrete-time stochastic process). *A stochastic process $\boldsymbol{y} = \{\boldsymbol{y}_t : t \in \mathbb{Z}\}$ is strictly stationary if for any $\{t_1, \dots, t_N\} \subset \mathbb{Z}$, the joint distribution of the random variables $\{\boldsymbol{y}(t_1 + \tau),\ \dots,\ \boldsymbol{y}(t_N + \tau)\}$ is independent of $\tau$.*

*It is weakly stationary, or wide-sense stationary, if it holds that*

$$\begin{aligned} \mathbb{E}[\boldsymbol{y}_t] &= \mu & \forall t &\in \mathbb{Z}, \\ \mathbb{E}[(\boldsymbol{y}_t - \mu)(\boldsymbol{y}_{t+\tau} - \mu)^\top] &= R_y(\tau) & \forall \tau &\in \mathbb{Z}. \end{aligned}$$

A weaker concept of stationarity, commonly used in system identification (specifically when LTI models are used), is quasi-stationarity. It allows for a common framework for stochastic and deterministic signals, and is used to imply that the signals satisfy regularity conditions (a form of ergodicity) used for the asymptotic analysis of the identification methods.

**Definition 2.1.3** (Quasi-stationary stochastic process [92, Section 2.5]). *A stochastic process $\boldsymbol{y} = \{\boldsymbol{y}_t : t \in \mathbb{Z}\}$ is quasi-stationary if*

$$\begin{aligned} \mathbb{E}[\boldsymbol{y}_t] &= \mu_t, & \|\mu_t\| &\leq C, & \forall t &\in \mathbb{Z}, \\ \mathbb{E}[\boldsymbol{y}_t \boldsymbol{y}_s^\top] &= R_y(t,s), & \|R_y(t,s)\| &\leq C, & \forall t, s &\in \mathbb{Z}, \\ \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} R_y(t, t-\tau) &= R_y(\tau), & & & \forall \tau &\in \mathbb{Z}, \end{aligned}$$

*in which the expectation operator is with respect to the distribution of the random component of the signal. If the signal is deterministic, then the expectation operator can be omitted.*

A special class of second-order stochastic processes is the class of white noise processes.

**Definition 2.1.4** (White noise). *A stochastic process $\boldsymbol{\zeta} = \{\boldsymbol{\zeta}_t : t \in \mathbb{Z}\}$ is white noise if $\mathbb{E}[\boldsymbol{\zeta}_t] = 0$, $\|\mathbb{E}[\boldsymbol{\zeta}_t \boldsymbol{\zeta}_t^\top]\| < \infty\ \ \forall t \in \mathbb{Z}$, and $\mathbb{E}[\boldsymbol{\zeta}_t \boldsymbol{\zeta}_s^\top] = 0\ \ \forall t \neq s$. In words, white noise is a sequence of uncorrelated random variables with zero mean and finite second order moments.*

This definition of white noise is “weak” and is used when the estimation method does not rely on the exact distribution of the process, but only on the first and second moments. In this case, the exact finite-dimensional distributions of the process are not specified. However, it is sometimes required to work with white noise which is a sequence of independent random variables; in this case, we speak of an independent process (or independent white noise). Furthermore, in some cases, it is assumed that the white noise is an independent and identically distributed (i.i.d.) process (following a Gaussian distribution for example).

In system identification, the disturbances (uncertain errors) are usually understood to come from two main sources. The first source is the imperfections of the sensing devices used to measure the outputs; this is known as measurement noise. The second source is the uncontrollable inputs (passing through the system) that affect the observed outputs; this is known as process disturbances. In the linear setting, the measurement noise is commonly assumed to be white noise, while the process disturbances are usually modeled as linearly filtered white noise. The inputs are normally assumed to be known deterministic signals or known realizations of stochastic processes; in either case, it is usually assumed that the input is quasi-stationary. Under some assumptions on the data-generation mechanism, it is possible to show that the output is also a quasi-stationary signal.

A mathematical model in this thesis is understood to be the rule that specifies the evolution of the signals through time. In other words, it is the deterministic structure underlying the stochastic observations. A relevant result that gives interesting insights regarding the structure of certain classes of stochastic processes is Wold’s decomposition introduced in [148] and its extension in [27].

### Wold’s decomposition

Consider a vector-valued discrete-time stochastic process $\boldsymbol{y} := \{\boldsymbol{y}_k\}_{k=-\infty}^{\infty} \subset \mathsf{L}_2^n(\Omega, \mathcal{F}, P_\theta)$. Here, the space $\mathsf{L}_2^n(\Omega, \mathcal{F}, P_\theta)$ is the Hilbert space of random variables, with zero mean and finite covariance, defined over $(\Omega, \mathcal{F}, P_\theta)$ and assuming values in $\mathbb{R}^n$ (see Appendix B). In what follows, we will be referring to this space simply as $\mathsf{L}_2^n$.

Define the Hilbert spaces

$$\boldsymbol{\mathcal{H}}_t := \overline{\mathbf{sp}}\{\boldsymbol{y}_s : s \leq t\}, \quad \forall t \in \mathbb{Z} \tag{2.1}$$

in which the symbol $\overline{\mathbf{sp}}\{\mathcal{S}\}$ is used to denote the closure of the span of $\mathcal{S} \subset \mathsf{L}_2^n$. Observe that we can always project $\boldsymbol{y}_t$ onto $\boldsymbol{\mathcal{H}}_{t-1}$ and define the difference

$$\boldsymbol{\varepsilon}_t := \boldsymbol{y}_t - \mathcal{P}_{\boldsymbol{\mathcal{H}}_{t-1}}[\boldsymbol{y}_t] \tag{2.2}$$

in which $\mathcal{P}$ denotes the projection operator (see Appendix B). The vector $\boldsymbol{\varepsilon}_t$ is known as the innovation in $\boldsymbol{y}_t$ and is orthogonal to $\boldsymbol{\mathcal{H}}_{t-1}$ by construction. The process $\boldsymbol{\varepsilon}$ is known as the “innovation process” of the process $\boldsymbol{y}$ (the name was suggested in [27]).

**Definition 2.1.5** (The (linear) innovation process). *The stochastic process defined by*

$$\boldsymbol{\varepsilon}_t := \boldsymbol{y}_t - \mathcal{P}_{\boldsymbol{\mathcal{H}}_{t-1}}[\boldsymbol{y}_t], \quad \boldsymbol{y}_t \in \mathsf{L}_2^n, \ \forall t \in \mathbb{Z}$$

*is called the innovation process of* $\boldsymbol{y}$*. If it holds that* $\mathbb{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\top] > 0 \ \forall t \in \mathbb{Z}$ *(i.e. the covariance matrix of* $\boldsymbol{\varepsilon}_t$ *is positive definite), then the process* $\boldsymbol{y}$ *is said to be full rank.*

By the definitions (2.1), (2.2), and the properties of $\mathcal{P}$, it holds that

$$\mathcal{P}_{\boldsymbol{\mathcal{H}}_{t-1}}[\boldsymbol{y}_t] = \mathcal{P}_{\{\boldsymbol{\varepsilon}_{t-1}\}}[\boldsymbol{y}_t] + \mathcal{P}_{\boldsymbol{\mathcal{H}}_{t-2}}[\boldsymbol{y}_t]$$

and therefore we may write

$$\boldsymbol{y}_t = \boldsymbol{\varepsilon}_t + \mathcal{P}_{\{\boldsymbol{\varepsilon}_{t-1}\}}[\boldsymbol{y}_t] + \mathcal{P}_{\boldsymbol{\mathcal{H}}_{t-2}}[\boldsymbol{y}_t], \quad t \in \mathbb{Z}.$$

Because $\boldsymbol{\varepsilon}$ has the same dimension as $\boldsymbol{y}$ and $\mathcal{P}$ is, by definition, a linear operator, it holds that

$$\mathcal{P}_{\{\boldsymbol{\varepsilon}_{t-1}\}}[\boldsymbol{y}_t] = h_1(t)\boldsymbol{\varepsilon}_{t-1}$$

in which $h_1(t)$ is a square matrix of real numbers. We can repeat the above projection $m$ times and write

$$\boldsymbol{y}_t = \sum_{k=0}^{m-1} h_k(t)\boldsymbol{\varepsilon}_{t-k} + \mathcal{P}_{\boldsymbol{\mathcal{H}}_{t-m}}[y_t] \tag{2.3}$$

in which $h_0(t) = I$. Now notice that the variance of the first term on the right hand side of (2.3) increases with $m$ but is bounded by the variance of $\boldsymbol{y}_t$. Due to the orthogonality of the two terms, the variance of the second term decreases with $m$ and is nonnegative. Therefore, asymptotically in $m$ we have

$$\begin{aligned} \boldsymbol{y}_t &= \sum_{k=0}^{\infty} h_k(t)\boldsymbol{\varepsilon}_{t-k} + \boldsymbol{y}_t^d \\ &= H_t(q)\boldsymbol{\varepsilon}_t + \boldsymbol{y}_t^d, \quad t \in \mathbb{Z} \end{aligned} \tag{2.4}$$

where $\{h_k(t)\Sigma_{t-k}\}_{k=0}^{\infty}$ is a square summable sequence. The representation of $\boldsymbol{y}$ in (2.4) is known as Wold's decomposition. Here, $\Sigma_{t-k}$ is a square root of $\mathbb{E}[\boldsymbol{\varepsilon}_{t-k}\boldsymbol{\varepsilon}_{t-k}^\top]$, $H_t(q) = \sum_{k=0}^{\infty} h_k(t)q^{-k}$ is a monic linear time-varying transfer operator (see Section 2.1.3) and $\boldsymbol{y}_t^d \in \boldsymbol{\mathcal{H}}_{t-m}$ for all $m \in \mathbb{N}$. This last observation means that $\boldsymbol{y}_t^d$ can be predicted perfectly given the history $\mathcal{H}_{t-m}$ (i.e., $\{y_k\}_{k=t-\infty}^{t-m}$), regardless of the value $m$. For this reason, the stochastic process $\boldsymbol{y}^d$ is called the linear deterministic part of $\boldsymbol{y}$. When this part is zero, the process $\boldsymbol{y}$ is known as a purely non-deterministic stochastic process and it can always be written as the output of a causal linear time-varying filter excited by white noise

$$\boldsymbol{y}_t = H_t(q)\boldsymbol{\varepsilon}_t = \boldsymbol{\varepsilon}_t + \sum_{k=1}^{\infty} h_k(t)\boldsymbol{\varepsilon}_{t-k}, \quad t \in \mathbb{Z}.$$

If the process $\boldsymbol{y}$ is weakly stationary, the coefficients of the filter $H$ are time independent and we may write

$$\boldsymbol{y}_t = H(q)\boldsymbol{\varepsilon}_t = \boldsymbol{\varepsilon}_t + \sum_{k=1}^{\infty} h_k \boldsymbol{\varepsilon}_{t-k}, \quad t \in \mathbb{Z}. \tag{2.5}$$

The decomposition into a deterministic and a purely non-deterministic part of a given second-order discrete-time stochastic process is the basis of time-domain linear prediction. It has a direct connection with the representation theorems for stationary stochastic processes ([7, Theorem 3.2], also see [120, 145]). Because Wold's decomposition is not based on the full distribution of the innovation process, it only captures the first and second moments of $\boldsymbol{y}$. This is sufficient whenever "linear" predictors are to be used.

We summarize Wold's decomposition of non-stationary stochastic processes in the following theorem.

**Theorem 2.1.6** (Extension of Wold's decomposition to non-stationary processes)**.** *For any given process $\boldsymbol{y}$ with finite second moments and mean function $m_t$, there is a uniquely determined decomposition*

$$\boldsymbol{y}_t - m_t = \boldsymbol{y}_t^r + \boldsymbol{y}_t^d \quad t \in \mathbb{Z}$$

*with the following properties:*

*(a) the processes $\boldsymbol{y}^r$ and $\boldsymbol{y}^d$ are orthogonal and $\boldsymbol{y}_t^r, \boldsymbol{y}_t^d \in \mathcal{H}_t \subset \mathsf{L}_2^n \ \ \forall t \in \mathbb{Z}$,*

*(b) the process $\boldsymbol{y}^d$ is linearly deterministic, i.e., $\boldsymbol{y}_t^d \in \mathcal{H}_{t-n} \subset \mathsf{L}_2^n \ \ \forall t, n \in \mathbb{N}$,*

*(c) the process $\boldsymbol{y}^r$ is purely non-deterministic and can always be expressed (linearly) in terms of the innovations of $\boldsymbol{y}$,*

$$\boldsymbol{y}_t^r = \sum_{k=0}^{\infty} h_k(t)\boldsymbol{\varepsilon}_{t-k}, \quad t \in \mathbb{Z}$$

*in which $\boldsymbol{\varepsilon}_t \subset \mathsf{L}_2^n$ is the innovation in $\boldsymbol{y}_t$ with a covariance matrix $\Lambda_t := \mathbb{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\top]$ satisfying*

$$\|\Lambda_t\| < \infty \ \ \forall t \in \mathbb{Z}, \quad \textit{and} \quad \mathbb{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j^\top] = 0 \ \ \forall k \neq j \in \mathbb{Z},$$

*where $\|\cdot\|$ is the squared Frobenius norm, such that*

$$\begin{aligned} &\sum_{k=0}^{\infty} h_k(t)\Lambda_{t-k}h_k^\top(t) \geq 0, \quad \sum_{k=0}^{\infty} \|h_k(t)\Lambda_{t-k}h_k^\top(t)\| < \infty \qquad \forall t \in \mathbb{Z}, \\ &h_n(t)\Lambda_{t-n} = \mathbb{E}[\boldsymbol{y}_t \boldsymbol{\varepsilon}_{t-n}^\top] = \mathbb{E}[\boldsymbol{y}_t^r \boldsymbol{\varepsilon}_{t-n}^\top] \quad \forall n \in \mathbb{N}_0, \ \textit{and} \\ &h_0(t)\Lambda_t = \Lambda_t = \Lambda_t h_0^\top(t), \quad \forall t \in \mathbb{Z}. \end{aligned} \tag{2.6}$$

*Furthermore, if the covariance matrix of $\boldsymbol{\varepsilon}_{t-n}$ $\forall n \in \mathbb{N}$, $\forall t \in \mathbb{Z}$ is full rank, the sequence $\{h_k(t) : k \in \mathbb{N}_0, \ t \in \mathbb{Z}\}$ is uniquely determined and the matrix $h_0(t) = I \ \forall t \in \mathbb{Z}$.*

*Proof.* The proof is due to Harold Cramér in [27] where he also discussed a similar decomposition for continuous-time processes. ■

The last part of Theorem 2.1.6 states that the second-order properties of a purely non-deterministic full rank process $\boldsymbol{y} \in \mathsf{L}_2^n$ correspond to the pair of sequences $(\{h_k(t) : k \in \mathbb{N}_0,\ t \in \mathbb{Z}\}, \{\Lambda_t : t \in \mathbb{Z}\})$. Once the second (the covariance sequence of the innovations) is given, the first is determined uniquely (see the second row of (2.6)). Observe that, in general, $\boldsymbol{y}_t^r$ can be written as the output of a time-varying filter with an impulse response sequence $\{h_k(t)\Sigma_t : k \in \mathbb{N}_0,\ t \in \mathbb{Z}\}$ and an input $\tilde{\varepsilon_t} = \Sigma_t^{-1}\varepsilon_t$ where $\Sigma_t$ is any sequence of positive definite square matrices such that $\{h_k(t)\Sigma_t\}$ is "square" summable. Therefore, we will be informally referring to the representation in Theorem 2.1.6 by Wold's decomposition even in cases where the covariance of $\varepsilon_t$ is not necessarily the innovation covariance.

In this thesis, we will be always assuming that the linear deterministic part $\boldsymbol{y}^d$ is identically zero. Furthermore, a property that we shall impose is the invertibility of the processes (w.r.t. the innovations). Observe that, by definition, $\boldsymbol{\varepsilon}_t \in \mathcal{H}_t$.

**Assumption 2.1.7.** *For every $t \in \mathbb{Z}$, there exists a uniformly exponentially decaying sequence $\{\tilde{h}_k(t) : k \in \mathbb{N}_0\}$ with $\tilde{h}_0(t) = I$ such that if we are given a realization $\{y_s : s \le t\}$, it is possible to write*

$$\varepsilon_t = \sum_{k=0}^{\infty} \tilde{h}_k(t)(y_{t-k} - m_{t-k}), \quad t \in \mathbb{Z}.$$

This assumption is used in the asymptotic analysis of prediction error methods (see Section 4.4); it ensures that the used predictors possess a required stability property. For the case of stationary processes, this property is linked to the spectral factorization theorem of strictly positive rational spectra (see [120]) and the invertibility of noise models in linear system identification (see [92, Section 3.2]).

**Example 2.1.1** (Wold's decomposition)**.** Consider the second-order discrete-time stationary stochastic process given by

$$\boldsymbol{y}_t = \boldsymbol{e}_t - 2\boldsymbol{e}_{t-1}, \quad t \in \mathbb{Z}, \tag{2.7}$$

where $\boldsymbol{e}_t$ is white noise with unit variance. Observe that we may write

$$0.5\boldsymbol{y}_t = 0.5\boldsymbol{e}_t - \boldsymbol{e}_{t-1} = (0.5q - 1)\boldsymbol{e}_{t-1}, \quad t \in \mathbb{Z}$$

in which $q$ is the forward shift operator (see [6]). Assuming zero initial conditions, the solution to this operator equation is given by

$$\boldsymbol{e}_{t-1} = -0.5 \sum_{k=0}^{\infty} 0.5^k \boldsymbol{y}_{t+k}, \tag{2.8}$$

showing that $\boldsymbol{e}_{t-1} \notin \mathcal{H}_{t-1}$. Therefore, (2.7) is not Wold's decomposition of $\boldsymbol{y}$.

To get Wold's decomposition, we need to write $\boldsymbol{y}_t$ in terms of the innovations $\{\boldsymbol{\varepsilon}_s\}_{s\leq t}$. Notice that, by redefining the white process in (2.7),

$$\begin{aligned}\boldsymbol{\varepsilon}_t - 0.5\boldsymbol{\varepsilon}_{t-1} = \boldsymbol{y}_t \;\Leftrightarrow\; \boldsymbol{\varepsilon}_t &= \sum_{k=0}^{\infty} 0.5^k \boldsymbol{y}_{t-k} \\ &= (\boldsymbol{e}_t - 2\boldsymbol{e}_{t-1}) + 0.5(\boldsymbol{e}_{t-1} - 2\boldsymbol{e}_{t-2}) + 0.5^2(\boldsymbol{e}_{t-2} - 2\boldsymbol{e}_{t-3}) + \ldots \\ &= \boldsymbol{e}_t - 3\sum_{k=1}^{\infty} 0.5^k \boldsymbol{e}_{t-k}\end{aligned}$$

and it follows that $\boldsymbol{\varepsilon}_t$ is the innovation process and Wold's decomposition of $\boldsymbol{y}$ is

$$\boldsymbol{y}_t = \boldsymbol{\varepsilon}_t - 0.5\boldsymbol{\varepsilon}_{t-1}, \quad \text{with} \quad \mathbf{var}(\boldsymbol{\varepsilon}_t) = 4 \;\; \forall t \in \mathbb{Z}.$$

It is obvious that Wold's decomposition is an incomplete representation that captures only the second-order properties of the process (compare to Problem 3T.4 in [92]).

### 2.1.2 Nonlinear Models

The model class considered in this thesis is the class of discrete-time causal dynamical models of the form

$$\boldsymbol{y}_t = f_t(\{u_k\}_{-\infty}^{t-1}, \{\boldsymbol{\zeta}_k\}_{-\infty}^{t}; \theta), \quad t \in \mathbb{Z} \tag{2.9}$$

in which $f_t(\cdot,\cdot;\theta)$ are general nonlinear maps parameterized by a finite-dimensional parameter vector $\theta \in \Theta$. The models map the history of the input process $\{u_k\}_{-\infty}^{t-1}$ and the history of the zero mean (usually stationary) process $\{\boldsymbol{\zeta}_k\}_{-\infty}^{t}$ to the current output $\boldsymbol{y}_t$. For each $t$, we assume that $u_t \in \mathbb{R}^{d_u}$ and $y_t \in \mathbb{R}^{d_y}$ for some $d_u, d_y \in \mathbb{N}$. This class is quite general and includes most of the commonly used models. So far, we only assume the finite-dimensional (sufficiently smooth) parameterization and the compactness of the set $\Theta$ such that the output process $\boldsymbol{y} \subset \mathsf{L}_2^{d_y}$.

Since in practice we only have access to finite input-output data records given in terms of two vectors

$$Y = [y_1^\top, \ldots, y_N^\top]^\top \in \mathbb{R}^{d_y N} \text{ and } U = [u_1^\top, \ldots, u_N^\top]^\top \in \mathbb{R}^{d_u N}$$

for some $N \in \mathbb{N}$, it is usually assumed that the history of the input $\{u_k\}_{-\infty}^{0}$ and the history of the disturbances $\{\boldsymbol{\zeta}_k\}_{-\infty}^{0}$ are identically zero. This amounts to a case with perfectly known initial conditions. When this assumption does not hold, it is sometimes possible to consider the unknown initial conditions as model parameters, or assume a random initial condition which is to be lumped to the process $\boldsymbol{\zeta}_t$ (observe that under stability assumptions, erroneous initial conditions do not influence the asymptotic properties of consistent estimators). This is summarized in the following definition.

**Definition 2.1.8.** *(Stochastic parametric nonlinear dynamical model) The stochastic nonlinear models are defined by the relations*

$$\boldsymbol{y}_t = f_t(\{u_k\}_{k=1}^{t-1}, \{\boldsymbol{\zeta}_k\}_{k=1}^{t}; \theta), \quad t = 1, 2, \ldots, N \in \mathbb{N}, \tag{2.10}$$

*in which $\theta \in \Theta \subset \mathbb{R}^d$ is a static parameter to be identified, and $\{\boldsymbol{\zeta}_k\}_{k=1}^{t}$ is a sequence of unobserved random vectors such that $\{\boldsymbol{y}_t : t = 1, \ldots, N\}$ is a subsequence of a second-order discrete-time stochastic process $\boldsymbol{y} = \{\boldsymbol{y}_t : t \in \mathbb{Z}\}$.*

The mappings $f_t$ between the inputs, the disturbances, and the outputs are general nonlinear maps; there are many ways to define such maps, either by using physical/semi-physical modeling, or by black-box modeling using general function expansions, see for example [134] or [92, Chapter 6] for an overview of possible nonlinear mapping based on basis functions. In this thesis, we are mainly interested in model classes for which the disturbance signal acts upon the output through a non-invertible nonlinear transformation, as presented in Section 1.2.2.

### Nonlinear state-space models

A model class of interest is the class of discrete-time nonlinear state-space models, which has been used in signal processing, systems and control theory with a wide range of applications. These models are defined by a set of first order difference equations and are often based on some physical insight (see [92, Section 4.3]). A causal stochastic discrete-time nonlinear state-space model is given by

$$\begin{aligned} \boldsymbol{x}_{t+1} &= h(\boldsymbol{x}_t, u_t, \boldsymbol{w}_t; \theta), & &\text{(the state equation)} \\ \boldsymbol{y}_t &= g(\boldsymbol{x}_t, \boldsymbol{v}_t; \theta), \quad t = 1, 2, \ldots, N. & &\text{(the output equation)} \end{aligned} \tag{2.11}$$

Such a model is often further restricted by assuming that the process disturbance $\boldsymbol{w}$ and the measurement noises $\boldsymbol{v}$ enter additively so that

$$\begin{aligned} \boldsymbol{x}_{t+1} &= h(\boldsymbol{x}_t, u_t; \theta) + \boldsymbol{w}_t \\ \boldsymbol{y}_t &= g(\boldsymbol{x}_t; \theta) + \boldsymbol{v}_t. \end{aligned} \tag{2.12}$$

The (latent/hidden) process $\boldsymbol{x}$ is known as the state process. Under some assumptions on the disturbance process and measurement noise (see below), the state process satisfies the Markov property. For brevity, we will suppress the dependence on $u_t$ in all the notations in this part.

**Definition 2.1.9.** *(The Markov property) A stochastic process $\boldsymbol{x} = \{\boldsymbol{x}_t : t \in \mathbb{Z}\}$ is said to have the Markov property if for any finite index set $\{t_i : t_i < t_{i+1}\} \subset \mathbb{Z}$, it holds that*

$$p(\boldsymbol{x}_{t_i} | x_{t_1}, \ldots x_{t_{i-1}}; \theta) = p(\boldsymbol{x}_{t_i} | x_{t_{i-1}}; \theta). \tag{2.13}$$

*In this case, the process $\boldsymbol{x}$ is said to be a Markov process.*

This property indicates that the information in the history of the process regarding $\boldsymbol{x}_t$ is summarized in $\boldsymbol{x}_{t-1}$. When looking at $\boldsymbol{x}_t$, the Markov property makes it possible to model the whole past of the process with a single initial condition $\boldsymbol{x}_0$. The uncertain infinite past $\{\boldsymbol{x}_{t-n} : n \in \mathbb{N}\}$ can then be modeled by letting

$$\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0; \theta) \tag{2.14}$$

which is known as the prior over $\boldsymbol{x}_0$.

We now show that, when both the measurements and the disturbance process are i.i.d. white noises that are mutually independent, the latent process $\{\boldsymbol{x}_t\}$ generated by the state-space model is a Markov process. The assumption means that $p(\boldsymbol{w}_t, \boldsymbol{w}_s; \theta) = p(\boldsymbol{w}_t; \theta)p(\boldsymbol{w}_s; \theta)$ for all $t \neq s$, and a similar relation holds for the noise process.

Assume that the sequence $\{x_k\}_{k=0}^t$ is given; that is $\boldsymbol{x}_k = x_k$ for $k = 0, \ldots, t$. According to the state equation (first row of (2.12)), it holds that

$$\boldsymbol{x}_{t+1} | \{x_k\}_{k=0}^t \overset{\mathrm{d}}{=} h(x_t, u_t; \theta) + \boldsymbol{w}_t | \{x_k\}_{k=0}^t \tag{2.15}$$

where we used $\overset{\mathrm{d}}{=}$ to denote equality in distribution, and the notation $\boldsymbol{\zeta}|\eta$ to denote the random variable $\boldsymbol{\zeta} \sim p(\boldsymbol{\zeta}|\boldsymbol{\eta} = \eta)$. Due to the independence assumption on the process disturbance, it then holds that

$$\boldsymbol{w}_t | \{x_k\}_{k=0}^t \overset{\mathrm{d}}{=} \boldsymbol{w}_t$$

and

$$\boldsymbol{x}_{t+1} | \{x_k\}_{k=0}^t \overset{\mathrm{d}}{=} h(x_t, u_t; \theta) + \boldsymbol{w}_t \overset{\mathrm{d}}{=} \boldsymbol{x}_{t+1} | x_t, \tag{2.16}$$

and therefore the state process is a Markov process if the disturbance process is independent. By conditioning on $\boldsymbol{x}_t$, the model can be rewritten as

$$\boldsymbol{x}_{t+1} \sim p(\boldsymbol{x}_{t+1} | x_t; \theta)$$

where sampling (simulating) according to $p(\boldsymbol{x}_{t+1}|x_t; \theta)$ is done in two steps. First, a sample $\boldsymbol{w}_t \sim p(\boldsymbol{w}_t; \theta)$ is generated; then the state equation is used to define $\boldsymbol{x}_{t+1} = h(x_t, u_t; \theta) + \boldsymbol{w}_t$. We also observe that the probability that $\boldsymbol{x}_{t+1} = x_{t+1}$ is given by

$$p(x_{t+1}|x_t; \theta) = p_{\boldsymbol{w}}(x_{t+1} - h(x_t, u_t; \theta); \theta).$$

in which the notation $p_{\boldsymbol{w}}(w; \theta)$ denotes the value $p(\boldsymbol{w}_t = w; \theta)$.

Moreover, conditioning on the state process, the outputs $\boldsymbol{y}_t$ become independent due to the independence assumption on the measurement noise, and it holds that

$$\boldsymbol{y}_t \sim p(\boldsymbol{y}_t | x_t; \theta).$$

Sampling according to $p(\boldsymbol{y}_t|x_t; \theta)$ is done in two steps. First, a sample $\boldsymbol{v}_t \sim p(\boldsymbol{v}_t; \theta)$ is generated; then the output equation is used to define $\boldsymbol{y}_t = g(x_t, u_t; \theta) + \boldsymbol{v}_t$. The probability that $\boldsymbol{y}_t = y_t$ is given by

$$p(y_t|x_t; \theta) = p_{\boldsymbol{v}}(y_t - g(x_t, u_t; \theta); \theta).$$

Finally, the joint PDF of the states up to time $t$ and the joint PDF of the outputs up to time $t$ are given by

$$\begin{aligned} p(\boldsymbol{X}_t;\theta) &= p(\boldsymbol{x}_0;\theta)\prod_{k=1}^{t-1} p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k;\theta), \\ p(\boldsymbol{Y}_t|\boldsymbol{X}_t;\theta) &= \prod_{k=1}^{t} p(\boldsymbol{y}_k|\boldsymbol{X}_t;\theta) = \prod_{k=1}^{t} p(\boldsymbol{y}_k|\boldsymbol{x}_k;\theta). \end{aligned} \tag{2.17}$$

where $\boldsymbol{X}_t := [\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t := [\boldsymbol{y}_1^\top, \dots, \boldsymbol{y}_t^\top]^\top$. These PDFs are used to construct the likelihood function of the model parameters.

### 2.1.3 Linear Models

For Linear Time-Invariant (LTI) models, the mapping between the inputs, the disturbances and the outputs is a linear time-independent map. A causal stochastic discrete-time LTI model is given by

$$\boldsymbol{y}_t = \sum_{k=1}^{\infty} g_k u_{t-k} + \sum_{k=0}^{\infty} h_k \boldsymbol{e}_{t-k}, \quad t \in \mathbb{Z}. \tag{2.18}$$

The model is therefore completely characterized by the impulse response sequences $\{g_k\}$ and $\{h_k\}$, together with all the finite-dimensional distributions of the process $\boldsymbol{e} = \{\boldsymbol{e}_k\}$. It is usually assumed that at least one input delay is present. This assumption holds for many practical discrete-time systems and is usually invoked when considering systems with feedback mechanisms. The process $\boldsymbol{e}$ represents possible measurement noise and/or internal process disturbances. Due to the linearity of the model, the contribution from all sources of randomness can be modeled with the process $\boldsymbol{e}$, which is usually assumed to be white noise. Observe that this assumption can be motivated by Wold's decomposition described in Theorem 2.1.6.

It is common to work with structures that allow the specification of the infinite impulse response sequences in terms of a finite-dimensional parameter.

#### Transfer operator models

Transfer operators are defined in terms of shift operators. It is common to use the backward shift operator $q^{-1}$ which is defined by the relation

$$q^{-1}u_k = u_{k-1}$$

where it is assumed that all signals are doubly infinite sequences (see [151, Chapter 7]). A shift-operator algebra allowing division by polynomials whose roots (strictly) inside the unit disc is usually used ([6, Section 2.6] and [92, Lemma 3.1]).

A transfer operator corresponding to an impulse response sequence $\{g_k\}$ is defined by

$$G(q, \{g_k\}) := \sum_{k=1}^{\infty} g_k q^{-k}.$$

This is an operator acting on the space of doubly infinite (inputs/disturbances) sequences. To be able to characterize models in terms of finite-dimensional parameters in the case of Single-Input Single-Output (SISO) models, we restrict the impulse response sequences to those that can be obtained as an expansion of rational functions and define

$$G(q,\theta) = \sum_{k=1}^{\infty} g_k(\theta) q^{-k} = \frac{B(q,\theta)}{F(q,\theta)}$$

in which $B(q,\theta)$ and $F(q,\theta)$ are polynomials in the shift operator whose coefficients are entries of $\theta$.

A SISO LTI model with a finite parameter can then be written as

$$\begin{aligned} \boldsymbol{y}_t &= \sum_{k=1}^{\infty} g_k(\theta) u_{t-k} + \sum_{k=0}^{\infty} h_k(\theta) \boldsymbol{e}_{t-k} \\ &= G(q,\theta) u_t + H(q,\theta) \boldsymbol{e}_t \\ &= \frac{B(q,\theta)}{F(q,\theta)} u_t + \frac{C(q,\theta)}{D(q,\theta)} \boldsymbol{e}_t, \qquad t \in \mathbb{Z}. \end{aligned} \tag{2.19}$$

The transfer operator $G$ acting on the input is usually known as the plant model, while the transfer operator $H$ acting on the stochastic signal is known as the noise model. The plant model assumes at least one input delay and it is usually assumed that $H$ is monic, that is $h_0(\theta) = 1$. This last assumption is not restrictive, since we can parameterize the variance of the white noise $\boldsymbol{e}_t$. A detailed description of possible choices for models with rational functions is given in [92].

For Multiple-Input Multiple-Output (MIMO) models, a finite parameterization can be achieved using matrix fraction descriptions, see [92, Appendix 4A] or [154].

Observe that, under the assumption that the history of the signals for all $t \le 0$ is known to be zero, the following vector equation holds:

$$\begin{aligned} \underbrace{\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{y}_3 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}}_{=\boldsymbol{Y}} &= \underbrace{\begin{bmatrix} 0 & 0 & \dots & 0 \\ g_1(\theta) & 0 & \dots & 0 \\ g_2(\theta) & g_1(\theta) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g_{N-1}(\theta) & g_{N-2}(\theta) & \dots & g_1(\theta) \end{bmatrix}}_{=:G(\theta)} \underbrace{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{bmatrix}}_{:=U} \\ &\quad + \underbrace{\begin{bmatrix} 1 & 0 & \dots & 0 \\ h_1(\theta) & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{N-1}(\theta) & h_{N-1}(\theta) & \dots & 1 \end{bmatrix}}_{=:H(\theta)} \underbrace{\begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_N \end{bmatrix}}_{=:\boldsymbol{E}}, \end{aligned} \tag{2.20}$$

that is,

$$\boldsymbol{Y} = G(\theta)U + H(\theta)\boldsymbol{E}. \tag{2.21}$$

Due to the assumption that $h_0(\theta) = 1$, the matrix $H(\theta)$ is a lower unitriangular (lower triangular with unit diagonal entires) matrix and therefore invertible. This means that we can write the vector $\boldsymbol{E}$ in terms of the observations $\boldsymbol{Y}$ and the inputs $U$ by inverting $H$;

$$\boldsymbol{E} = H^{-1}(\theta)(\boldsymbol{Y} - G(\theta)U). \tag{2.22}$$

Also observe that whenever $\boldsymbol{E}$ has a zero mean and a covariance matrix $\Lambda_N$, the vector $\boldsymbol{Y}$ has a mean value $\mu_Y(U;\theta) = G(\theta)U$ and a covariance $\Sigma_Y(\theta) = H(\theta)\Lambda_N H^\top(\theta)$ regardless of the distribution of $\boldsymbol{E}$. Hence, the first and second moments of $\boldsymbol{Y}$ are always available in closed-form.

**Linear state-space models**

Discrete-time linear state-space models are a special case of nonlinear state-space models; they describe the relation between the signals using a set of first order linear difference equations via the state process $\boldsymbol{x}$. They are defined by

$$\begin{aligned} \boldsymbol{x}_{t+1} &= A(\theta)\boldsymbol{x}_t + B(\theta)u_t + \boldsymbol{w}_t, \quad t \in \mathbb{N}_0 \\ \boldsymbol{y}_t &= C(\theta)\boldsymbol{x}_t + \boldsymbol{v}_t. \end{aligned} \tag{2.23}$$

in addition to the PDFs of $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$. The matrices $A, B,$ and $C$ are parameterized matrices with appropriate dimensions. Linear state-space representations are not unique; they depend on the choice of the coordinates for the state vector. However, there are several canonical forms that link state-space representations to rational transfer operators (see [68]).

Observe that we can write

$$\boldsymbol{y}_t = C(\theta)A^t(\theta)\boldsymbol{x}_0 + \sum_{k=0}^{t-1} C(\theta)A^k(\theta)B(\theta)u_{t-k-1} + \sum_{k=0}^{t-1} C(\theta)A^k(\theta)\boldsymbol{w}_{t-k-1} + \boldsymbol{v}_t \tag{2.24}$$

for $t = 1, 2, \ldots, N$. The first term on the right hand side is due to the initial state $\boldsymbol{x}_0$. Given $x_0$, both the history of the input $\{u_k\}_{k=-\infty}^{-1}$ and the history of the process disturbance $\{w_k\}_{k=-\infty}^{-1}$ are not needed to compute the value of the output at any $t \geq 0$. Assuming that the initial state $\boldsymbol{x}_0 = 0$, it holds that

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{y}_3 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}}_{=:\boldsymbol{Y}} = \underbrace{\begin{bmatrix} 0 & 0 & \dots & 0 \\ C(\theta)B(\theta) & 0 & \dots & 0 \\ C(\theta)A(\theta)B(\theta) & C(\theta)B(\theta) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C(\theta)A^{N-2}(\theta)B(\theta) & C(\theta)A^{N-3}(\theta)B(\theta) & \dots & C(\theta)B(\theta) \end{bmatrix}}_{=:G(\theta)} \underbrace{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{bmatrix}}_{=:U}$$

$$+ \underbrace{\begin{bmatrix} C(\theta) & 0 & \dots & 0 \\ C(\theta)A(\theta) & C(\theta) & \dots & 0 \\ C(\theta)A^2(\theta) & C(\theta)A(\theta) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C(\theta)A^{N-1}(\theta) & C(\theta)A^{N-2}(\theta) & \dots & C(\theta) \end{bmatrix}}_{=:F(\theta)} \underbrace{\begin{bmatrix} \boldsymbol{w}_0 \\ \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_{N-1} \end{bmatrix}}_{=:\boldsymbol{W}} + \underbrace{\begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \boldsymbol{v}_3 \\ \vdots \\ \boldsymbol{v}_N \end{bmatrix}}_{=:\boldsymbol{V}},$$

that is

$$\boldsymbol{Y} = G(\theta)U + F(\theta)\boldsymbol{W} + \boldsymbol{V}. \tag{2.25}$$

Unlike (2.21), the model in (2.25) is not invertible as it is, i.e., the outputs $\boldsymbol{Y}$ and the input $U$ cannot be used to recover $\boldsymbol{W}$ and $\boldsymbol{V}$. However, due to the linearity of the model, it can be transformed into an invertible form.

Assume that $\boldsymbol{W}$ and $\boldsymbol{V}$ are uncorrelated with zero mean and covariance matrices $\Sigma_W(\theta)$ and $\Sigma_V(\theta)$ respectively. Then, the first and second moments of $\boldsymbol{Y}$ are preserved if the random vector $F(\theta)\boldsymbol{W} + \boldsymbol{V}$ in (2.25) is replaced by

$$(F(\theta)\Sigma_W(\theta)F^\top(\theta) + \Sigma_V(\theta))^{\frac{1}{2}}\boldsymbol{E} \tag{2.26}$$

in which $\boldsymbol{E}$ is a zero mean vector with the identity covariance matrix (for example, a standard Gaussian random vector), and $(\cdot)^{\frac{1}{2}}$ denotes one of the possible matrix square roots (for example, the Cholesky square root; see [61, Section 6.4]). By defining the lower triangular matrix

$$H(\theta) := (F(\theta)\Sigma_W(\theta)F^\top(\theta) + \Sigma_V(\theta))^{\frac{1}{2}},$$

the model can be written in the invertible form

$$\boldsymbol{Y} = G(\theta)U + H(\theta)\boldsymbol{E} \tag{2.27}$$

which has the same form as (2.21). Observe that here, $G(\theta)$ and $H(\theta)$ are jointly parameterized by $\theta$ (via $A(\theta)$), and that $H(\theta)$ is triangular but not unitriangular. Moreover, note that the representation in (2.27) only captures the first and second moments of $\boldsymbol{Y}$.

## 2.2 The True System

It is always true that the assumed mathematical model does not exactly describe the underlying "real-life" system. It is therefore always seen as a mere approximation. However, for the purpose of analysis, it is convenient to assume the existence of a true system in order to understand the behavior of different identification methods. For this purpose, we will make the following assumption.

**Assumption 2.2.1** (True system)**.** *The observed data follow a known mathematical rule defined by a true parameter $\theta^\circ \in \Theta \subset \mathbb{R}^d$ such that*

$$\boldsymbol{y}_t = f_t(\{u_k\}_{k=1}^{t-1}, \{\zeta_k\}_{k=1}^{t}; \theta^\circ) \in \mathsf{L}_2^n, \quad t = 1, 2, \dots, N, \tag{2.28}$$

*for some known functions $f_t$, known inputs $\{u_k\}$, and unobserved disturbances $\{\zeta_k\}$.*

## 2.3 Estimators and Their Qualitative Properties

Estimators are defined as (measurable) functions of the observations. Let $\Theta$ be the parameter space for a parametric family of models corresponding to $P_\theta$, and let $\mathbb{E}[\cdot;\theta]$ denote the expectation with respect to the model with a parameter $\theta \in \Theta$. Define the data set

$$\boldsymbol{D}_t := \{(\boldsymbol{y}_k, u_k) : k = 1, \dots t\}, \; 1 \le t \le N, \tag{2.29}$$

that contains the outputs and inputs up to time $t$. We are interested in point estimators

$$\boldsymbol{D}_N \mapsto \hat{\boldsymbol{\theta}} := \hat{\theta}(\boldsymbol{D}_N) \in \Theta$$

that map the data to a point in the parameter space. We shall assume that the data is generated by a model governed by a parameter $\theta^\circ \in \Theta$ according to Assumption 2.2.1. It is important then to understand the relation between the process $\{\hat{\theta}(\boldsymbol{D}_N) : N \in \mathbb{N}\}$ and the parameter $\theta^\circ$. Because $\theta^\circ$ is unknown, any property of the considered estimators is desired to hold equally for all possible $\theta^\circ \in \Theta$.

We will now discuss some desired properties of point estimators. For the following definitions, we assume that all estimators have finite first and second moments.

**Definition 2.3.1** (Unbiased estimators)**.** *An estimator $\hat{\boldsymbol{\theta}}$ of a parameter $\theta$ is an unbiased estimator if*

$$\mathbb{E}[\hat{\theta}(\boldsymbol{D}_N); \theta] = \theta, \quad \forall \theta \in \Theta, \; \forall N \in \mathbb{N}.$$

*It is an asymptotically unbiased*[1] *estimator if*

[1] Note that some authors in the statistical inference literature define asymptotically unbiased estimators as those estimators with the property that $\mathbb{E}[\lim_{N\to\infty} f(N)\|\hat{\theta}(\boldsymbol{D}_N) - \theta\|; \theta] = 0 \; \forall \theta \in \Theta$ in which $f(N)$ is a normalization sequence (see [82, Chapter 6]). This definition (without an additional condition of uniform integrability) does not coincide with Definition 2.3.1.

$$\mathbb{E}[\hat{\theta}(\boldsymbol{D}_N);\theta] \to \theta \ \ as\ N \to \infty \quad \forall \theta \in \Theta.$$

*Otherwise, it is an asymptotically biased*[2] *estimator.*

**Definition 2.3.2** (Consistency). *An estimator $\hat{\boldsymbol{\theta}}$ of a parameter $\theta$ is a consistent estimator if*

$$\hat{\theta}(\boldsymbol{D}_N) \xrightarrow{a.s.} \theta \quad as\ N \to \infty \quad \forall \theta \in \Theta.$$

*in which $\xrightarrow{a.s.}$ denotes almost sure convergence (see [23, Chapter 4]).*

**Definition 2.3.3** ((Statistical) efficiency). *A consistent estimator $\hat{\boldsymbol{\theta}}$ of a parameter $\theta$ is a minimum variance/optimal or (statistically) efficient estimator if, for any other consistent estimator $\tilde{\boldsymbol{\theta}}$ of $\theta$, it holds that*

$$\mathbb{E}[(\tilde{\theta}(\boldsymbol{D}_N) - \theta)(\tilde{\theta}(\boldsymbol{D}_N) - \theta)^\top;\theta] \geq \mathbb{E}[(\hat{\theta}(\boldsymbol{D}_N) - \theta)(\hat{\theta}(\boldsymbol{D}_N) - \theta)^\top;\theta] \ \textit{for every } N.$$

*The estimator $\hat{\boldsymbol{\theta}}$ is said to be asymptotically efficient if, for all other estimators $\tilde{\boldsymbol{\theta}}$ of $\theta$, it holds that*

$$\lim_{N\to\infty} N\,\mathbb{E}[(\tilde{\theta}(\boldsymbol{D}_N) - \theta)(\tilde{\theta}(\boldsymbol{D}_N) - \theta)^\top;\theta] \geq \lim_{N\to\infty} N\,\mathbb{E}[(\hat{\theta}(\boldsymbol{D}_N) - \theta)(\hat{\theta}(\boldsymbol{D}_N) - \theta)^\top;\theta].$$

*Otherwise, it is a suboptimal estimator.*

**Definition 2.3.4** (Asymptotic normality). *An estimator $\hat{\boldsymbol{\theta}}$ is asymptotically normal about $\theta$ if there exists a sequence of matrices $\{P_N\}$ such that $P_N \succ \delta I$ for some $\delta \in \mathbb{R}_+$, all sufficiently large $N \in \mathbb{N}$, and*

$$\sqrt{N} P_N^{-1/2}(\hat{\theta}(\boldsymbol{D}_N) - \theta) \rightsquigarrow \mathcal{N}(0, I) \quad as \quad N \to \infty.$$

*Furthermore, if $P_N \to P \succ 0$ as $N \to \infty$, we say that $\hat{\boldsymbol{\theta}}$ is asymptotically normal about $\theta$ with asymptotic covariance matrix $P$. The symbol $\rightsquigarrow$ denotes convergence in distribution of random variables (see [23, Chapter 4]).*

The consistency of estimators is the weakest asymptotic property that should be required from a “good” estimator. Consistent estimators can be used to construct efficient estimators (see for example [84, Theorem 7.3.3], [138, Problem 7.12] or [143]). The asymptotic normality of estimators is important for the construction of approximate confidence intervals.

## 2.4 Estimation Methods

In this thesis, we mainly focus on two (frequentist/Fisherian, see [35]) estimation methods: the Maximum Likelihood Method and the Prediction Error Methods. In the following subsections, we summarize the definition and some properties of each method.

[2] In the text, we simply say that the estimator is biased to mean that it is asymptotically biased.

### 2.4.1 The Maximum Likelihood Method

The Maximum Likelihood (ML) method is a statistical inference method based on the likelihood principle. The principle essentially states that all the information/evidence obtained from an experiment about $\theta$ is contained in the likelihood function of $\theta$ (see Definition 2.4.1 below) for the given data. The idea has been used informally as early as the 1700s, by statisticians such as Carl Friedrich Gauss, Pierre-Simon Laplace, Thorvald N. Thiele, and Francis Ysidro Edgeworth [140]. However, it is usually attributed to Sir Ronald A. Fisher (1890–1962) who promoted the method and contributed to the theoretical analysis of its properties, see [40, 41]. Maximum Likelihood Estimators (MLE) are examples of the more general class of extremum estimators (see [1, Chapter 4]) – a general class of estimators based on the maximization of an objective function of the data.

**Definition 2.4.1** (Likelihood function)**.** *The likelihood function of the parameter $\theta$ for a given realization $Y$ of the model outputs and known inputs $U$ is defined by the nonnegative real function*

$$p(Y|U;\cdot) : \Theta \to \mathbb{R}_+.$$

*In addition, the function*

$$\log p(Y|U;\cdot) : \Theta \to \mathbb{R}_+$$

*is referred to as the log-likelihood function.*

Because, in what follows, it will be always assumed that $U$ is known, we will remove it from the notation and refer to the likelihood function of $\theta$ simply by $p(Y;\theta)$ seen as a function of $\theta$.

**Definition 2.4.2** (Maximum Likelihood Estimator)**.** *The random variable*

$$\hat{\theta}(\boldsymbol{D}_N) \;\; := \arg\max_{\theta\in\Theta} p(\boldsymbol{Y},\theta) \tag{2.30}$$

*(when it exists) is the Maximum Likelihood Estimator. Given a realization $Y$, any element of the set of global minimizers $\hat{\theta}(D_N)$ is called a maximum likelihood estimate.*

From the above definitions, it is obvious that the ML method requires a full probabilistic model.

Next, assuming that the likelihood function is differentiable with respect to $\theta$, we define the score function of the model. It gives an indication of how sensitive the likelihood function is to variations in $\theta$.

**Definition 2.4.3** (Score function)**.** *The random vector $\nabla_\theta \log p(\boldsymbol{Y};\theta)$ with coordinates $\frac{\partial}{\partial\theta_i}\log p(\boldsymbol{Y};\theta)$ is called the score function. A realization of the score function corresponding to $Y$ is equal to the gradient of the logarithm of the likelihood function of $\theta$ evaluated at $Y$.*

The covariance of the score function is known as the Fisher information matrix, and, under some regularity conditions, is equivalently given by

$$-\mathbb{E}\left[\nabla_\theta^2 \log p(\boldsymbol{Y};\theta);\theta\right]\Bigg|_{\theta=\theta^\circ}$$

where $\nabla_\theta^2 \log p(\boldsymbol{Y};\theta)$ denotes the Hessian matrix of the log-likelihood function of $\theta$. The inverse of the Fisher information matrix gives the Cramér-Rao lower bound on the Mean-Square Error (MSE) matrix in the class of unbiased estimators (see [26, 113]). It also holds, under some regularity conditions as shown in [79] or [8], that the inverse of

$$\mathcal{I}_F(\theta^\circ) = \lim_{N\to\infty} \frac{1}{N}\mathbf{cov}\left(\nabla_\theta \log p(\boldsymbol{Y};\theta), \nabla_\theta \log p(\boldsymbol{Y};\theta)\right)\Bigg|_{\theta=\theta^\circ}$$

gives a lower bound on the "asymptotic MSE" of most estimators, except for parameters in a set of Lebesgue measure zero. Under some weak conditions, see for example [5] or [60], the MLE of dynamical models was shown to be consistent, asymptotically normal, and asymptotically efficient (i.e. with covariance matrix $\mathcal{I}_F^{-1}(\theta^\circ)$). It is because of these appealing theoretical properties that the MLE has been used in many scientific fields, including system identification. However, it is worth mentioning that the ML method is justified asymptotically and has no finite sample guarantees. Furthermore, depending on the model, the MLE might not be well defined and can be inconsistent as shown for example in [80].

The method of ML was first introduced in the system identification community in [5], where it was used to estimate the parameters of linear dynamical models. The next example defines the MLE of a first order LTI state-space model.

**Example 2.4.1** (MLE of an LTI dynamical system)**.** Assume that the outputs $Y$ were collected according to the state-space model

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \alpha\boldsymbol{x}_k + \beta u_t + \boldsymbol{w}_t, & \boldsymbol{w}_t &\sim p(\boldsymbol{w}_t), & x_0 &= 0, \\ \boldsymbol{y}_t &= \boldsymbol{x}_t + \boldsymbol{v}_t, & \boldsymbol{v}_t &\sim p(\boldsymbol{v}_t), & t &= 1,\ldots,N, \end{aligned} \tag{2.31}$$

with $\alpha = \alpha_\circ$ and $\beta = \beta_\circ$. Furthermore, assume that the processes $\boldsymbol{w}$ and $\boldsymbol{v}$ are mutually independent i.i.d. white noise processes. Let $\theta := [\alpha \;\; \beta]^\top$, and observe that the model can be written in the vector form as in (2.25) for some parameter dependent matrices $G(\theta)$ and $F(\theta)$.

The associated family of PDFs of $\boldsymbol{Y}$ is

$$\left\{p(\boldsymbol{Y};\theta) : p(\boldsymbol{Y};\theta) = \int p_V(\boldsymbol{Y} - G(\theta)U - F(\theta)W)p(W)\,\mathrm{d}W,\ \theta\in\Theta\right\}$$

in which $p(\boldsymbol{V}) = \prod_{t=1}^N p(\boldsymbol{v}_t)$ and $p(\boldsymbol{W}) = \prod_{t=1}^N p(\boldsymbol{w}_t)$. If these distributions are not Gaussian, the above integral might not have a closed form solution. However, in cases where $\boldsymbol{w}$ and $\boldsymbol{v}$ are Gaussian processes with zero mean and covariances $\sigma_w^2$ and $\sigma_v^2$ respectively, the process $\boldsymbol{y}$ is Gaussian and the PDF of $\boldsymbol{Y}$ is

$$\{\, p(\boldsymbol{Y};\theta) : p(\boldsymbol{Y};\theta) = \mathcal{N}(\mu(U;\theta), \Sigma(\theta)),\ \theta \in \Theta \,\},$$

in which $\mu(U;\theta) = G(\theta)U$, and $\Sigma(\theta) = \sigma_w^2 F(\theta)F(\theta)^\top + \sigma_v^2 I$. The MLE is therefore given by

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\max_{\theta\in\Theta}\ p(\boldsymbol{Y},\theta)\\ &= \arg\max_{\theta\in\Theta}\ \frac{1}{(2\pi)^{\frac{N}{2}}\sqrt{\det\Sigma(\theta)}}\exp\left(-0.5(\boldsymbol{Y}-\mu(U,\theta))^\top\Sigma^{-1}(\theta)(\boldsymbol{Y}-\mu(U,\theta))\right).\end{aligned}$$

### 2.4.2 Prediction Error Methods

Prediction Error Methods (PEMs) are a large family of parameter estimation methods considered to be the corner stone of system identification (see [19, 92, 138]). They offer solutions to a wide range of problems and are supported by a solid underlying theory. Besides other advantages, they have both deterministic (see [91] or [92, Problem 8T.1]) as well as stochastic motivations, making them attractive for a wide range of applications. However, since we are concerned with stochastic models, we will only consider a stochastic framework here.

In this thesis, we will consider an estimation approach based on prediction error minimization. The main idea of this approach goes as follows. First, the parameterized model is written in terms of a parameterized predictor; that is a function of known inputs and previous observed outputs that "predicts" future outputs (see (2.32)). Then, the distance between the predicted output (the output of the predictor) and the observed output of the system is defined according to some metric. Such a metric, a positive scalar-valued function, can itself be parameterized either independently or by the model parameter vector itself. Finally, the method seeks the parameter that minimizes the metric over some predefined compact set $\Theta$.

If the predictor and the metric are selected according to the exact probabilistic nature of the data, the PEM coincides with the ML method (as explained below). Therefore, the PEMs can be seen as a generalization of the ML method and a specific instance of extremum estimators.

Let us define the parameterized one-step ahead predictor by the function

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \psi(\boldsymbol{D}_{t-1}, t; \theta), \quad t = 1, \dots, N, \quad \theta \in \Theta, \tag{2.32}$$

where we assume at least one input delay in the model. In practice, such a function is not always given as an explicit function of the data. It can be a function of a filtered version of the previous inputs and outputs (those before time $t$), or can even be constructed by running a Monte Carlo simulation as we suggest in this thesis (see Chapter 4). It is important to note that the stochastic assumptions used to determine the predictor function (2.32) do not have to be related to the exact full probabilistic structure of $\boldsymbol{D}_{t-1}$, which is an indication of the generality of the method. It is the representation of the predictor function given in (2.32) that matters. The importance of all the stochastic assumptions of the (true) model, the initial conditions, etc. is limited by the way they influence the definition of the predictor function. In other words, all the probabilistic assumptions that can be used to form a good predictor are immaterial, and the only important factor is the way these assumptions are reflected in the predictor. This remark was stressed in the contribution of Ljung in [90] and later in the book [92]. Also see [138, Sections 7.1-7.3, and Problem 7.3].

Once the predictor is defined, the estimation of the parameter $\theta$ is a matter of solving an optimization problem.

**Definition 2.4.4** (Prediction Error Methods estimator). *Given a predictor function $\psi$ and a nonnegative scalar-valued function $\ell$, the random variable*

$$\begin{aligned} \hat{\theta}(\boldsymbol{D}_N) := \min_{\theta\in\Theta} \quad & \sum_{t=1}^{N} \ell(\boldsymbol{e}_t(\theta), t; \theta) \\ \textit{such that} \quad & \boldsymbol{e}_t(\theta) = \boldsymbol{y}_t - \psi(\boldsymbol{D}_{t-1}, t; \theta), \quad \forall t = 1, \dots, N, \end{aligned} \tag{2.33}$$

*(when it exists) is the prediction error method estimator. Given a realization $D_N$, any element of the set of global minimizers $\hat{\theta}(D_N)$ is called a PEM estimate.*

The process $\boldsymbol{e}(\theta)$ is known as the prediction error process. Different choices for $\ell$ and the predictor function $\psi$ lead to different instances within the family of PEMs. A usual choice for $\ell$ is the quadratic Euclidean norm, which is time- and $\theta$-independent. With such a choice, the PEM problem (2.33) is an unweighted (nonlinear) least-squares problem. As we will show below, in some cases, the choice of $\ell$ is critical for the performance of the estimation method.

To summarize, for a given model and a data set, the two components that define an instance of the PEMs are

1. a parameterized predictor (2.32) for future outputs (function of the past data) – it can be defined according to the parameterized stochastic model with full or partial probabilistic assumptions, or it can be postulated directly.

2. a measure of distance, $\ell(\cdot; \theta)$, defined over the space of outputs $y$.

A question that arises naturally is: what are the best choices for these two components? The answer is given in the following.

### The best predictor

To define the best predictor for the outputs of a dynamical model, we first need to choose a criterion. In a stochastic framework, a commonly used natural measure (among other possibilities) is the MSE

$$\mathbb{E}\left[\|\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta)\|_2^2; \theta^\circ\right]. \tag{2.34}$$

It is not difficult to show that the optimal predictor in the sense of minimizing the MSE is given by the conditional expectation

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \mathbb{E}[\boldsymbol{y}_t|\boldsymbol{D}_{t-1};\theta]. \tag{2.35}$$

Such a predictor has an interesting geometrical interpretation in the Hilbert space of random variables $\mathsf{L}_2^n$. Let $\zeta$ denote an arbitrary measurable function of $\boldsymbol{D}_{t-1}$ representing candidate predictors. Then, using the tower property of expectations [23, Theorem 9.1.5], it holds that

$$\begin{aligned} &\mathbb{E}\big[\boldsymbol{y}_t^\top \zeta(\boldsymbol{D}_{t-1})\big] = \mathbb{E}\big[\mathbb{E}[\boldsymbol{y}_t^\top \zeta(\boldsymbol{D}_{t-1})|\boldsymbol{D}_{t-1}]\big] = \mathbb{E}\big[\mathbb{E}[\boldsymbol{y}_t|\boldsymbol{D}_{t-1}]^\top \zeta(\boldsymbol{D}_{t-1})\big] \\ \Rightarrow\ &\mathbb{E}\big[\ (\boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t|\boldsymbol{D}_{t-1}])^\top\ \zeta(\boldsymbol{D}_{t-1})\big] = 0 \end{aligned}$$

which shows that the error $\boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t|\boldsymbol{D}_{t-1}]$ is uncorrelated with (orthogonal to) any measurable function of $\boldsymbol{D}_{t-1}$. Thus, the error provides no "information" that can be used to better guess $\boldsymbol{y}_t$.

Observe that the optimal predictor relies on the probability distribution of the stochastic process $\boldsymbol{y}$ through its finite-dimensional distributions. In general, the computation of the optimal predictor involves a multidimensional integral (see (1.6) and (1.7)), which can be computed analytically only in very few cases. This is where the power of the PEM is truly apparent: it does not need an "optimal" predictor in order to construct consistent estimators.

### Kinship to the Maximum Likelihood Method

Let us now consider a narrow view of the PEMs ("narrow" because it restricts the choices of the predictor function and the function $\ell$ according to the exact (true) statistical properties of the data). Assume that the data is generated according to the data-generation mechanism

$$\boldsymbol{y}_t = \psi(\boldsymbol{D}_{t-1}, t; \theta) + \boldsymbol{e}_t, \quad t = 1, \dots, N, \tag{2.36}$$

in which $\boldsymbol{e}$ is an independent zero mean process, with a PDF $p(\boldsymbol{e}_t;\theta)$. The joint PDF of the observed outputs is easy to compute using the assumptions on $\boldsymbol{e}$;

$$p(\boldsymbol{Y};\theta) = \prod_{t=1}^{N} p_{e_t}(\boldsymbol{y}_t - \psi(\boldsymbol{D}_{t-1}, t; \theta); \theta).$$

Solving the ML estimation problem is equivalent to solving the minimization problem

$$\min_{\theta\in\Theta}\ -\log(p(\boldsymbol{Y};\theta)) = \min_{\theta\in\Theta}\ -\sum_{t=1}^{N} \log(p_{e_t}(\boldsymbol{y}_t - \psi(\boldsymbol{D}_{t-1}, t; \theta); \theta)).$$

To define a PEM estimator, let us use the model structure and parameterization as the one used for the true data (2.36). In this case, the optimal predictor (2.35) is easy to compute and is given by

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \psi(\boldsymbol{D}_{t-1}, t; \theta), \text{ and}$$
$$\boldsymbol{e}_t(\theta) = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta) \text{ with a PDF } p(\boldsymbol{e}_t; \theta).$$

Observe that we used the same symbol $\boldsymbol{e}_t$ for the prediction error process; it holds that $\boldsymbol{e}_t(\theta^\circ) = \boldsymbol{e}_t$. Therefore, if we use the optimal predictor and choose

$$\ell(\boldsymbol{e}_t, t; \theta) := -\log(p(\boldsymbol{e}_t; \theta)), \tag{2.37}$$

the solution of the PEM problem (2.33) coincides with the solution of the ML problem (2.30).

In the next example, we show how to construct the PEM problem for a stochastic LTI transfer operator model (2.19) parameterized by $\theta$. For details, see [92] or [138].

**Example 2.4.2** (PEM for a stochastic LTI transfer operator model)**.** Let us consider a parameterized LTI model in terms of a transfer operator

$$\boldsymbol{y}_t = G(q, \theta) u_t + H(q, \theta) \boldsymbol{e}_t, \quad t \in \mathbb{Z}, \tag{2.38}$$

in which the noise model $H(q, \theta)$ is monic and inversely stable and $\boldsymbol{e}$ is a martingale difference process, i.e., $\mathbb{E}[\boldsymbol{e}_t] = \mathbb{E}[\boldsymbol{e}_t | \{\boldsymbol{y}_k\}_{k<t}] = 0$.

It is then possible to construct an optimal one-step ahead predictor by inverting the noise model $H$. Observe that we may write

$$\boldsymbol{y}_t = [I - H^{-1}(q, \theta)] \boldsymbol{y}_t + H^{-1}(q, \theta) G(q, \theta) u_t + \boldsymbol{e}_t.$$

and therefore

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) := \psi(\boldsymbol{D}_{t-1}, \theta) = [I - H^{-1}(q, \theta)] \boldsymbol{y}_t + H^{-1}(q, \theta) G(q, \theta) u_t \tag{2.39}$$

is the optimal linear predictor, assuming that the data has been generated according to a mechanism like (2.38) (notice that the model does not specify the exact full distribution of $\boldsymbol{y}$). Observe that the optimal predictor is linear in both the known inputs and the previous outputs. For the stability of the predictor, the parameter $\theta$ must be constrained to the subset for which the two filters defining $\psi$ are stable.

Under the assumption that the histories $\{\boldsymbol{e}_k\}_{k\leq 0}$ and $\{u_k\}_{k\leq 0}$ are known to be identically zero, the outputs vector $\boldsymbol{Y}$ is given in a vector form as shown in (2.21). Let $\hat{\boldsymbol{Y}}(\theta) := \begin{bmatrix} \hat{\boldsymbol{y}}_{1|0}^\top(\theta) & \hat{\boldsymbol{y}}_{2|1}^\top(\theta) & \dots & \hat{\boldsymbol{y}}_{N|N-1}^\top(\theta) \end{bmatrix}^\top$, that is a column vector of stacked one-step ahead predictors. It then holds that

$$\hat{\boldsymbol{Y}}(\theta) = (I - H^{-1}(\theta)) \boldsymbol{Y} + H^{-1}(\theta) G(\theta) U \tag{2.40}$$

in which the matrix $G(\theta)$ and the vector $U$ are defined in (2.20) and

$$H^{-1}(\theta) = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \tilde{h}_1(\theta) & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{h}_{N-1}(\theta) & \tilde{h}_{N-1}(\theta) & \dots & 1 \end{bmatrix} \tag{2.41}$$

where $\{\tilde{h}_k(\theta)\}$ is the impulse response of the filter $H^{-1}(q,\theta)$ (it is exponentially decaying if the noise model is rational). Observe that $H^{-1}(\theta)$ coincides with the inverse of the matrix $H(\theta)$ defined in (2.20).

Another (suboptimal) choice for the one-step ahead predictor is

$$\hat{y}_{t|t-1}(\theta) = G(q,\theta)u_t \tag{2.42}$$

which ignores the probabilistic properties of the second term in (2.38) by assuming $H(q,\theta) = 1$. In this case, the predictor simulates the known input using the model $G(q,\theta)$, and the mean of the process $\boldsymbol{y}$ is used as a predictor. The predictor in (2.42) is known as the Output-Error (OE) predictor and is given in vector form by

$$\hat{Y}(\theta) = G(\theta)U. \tag{2.43}$$

The main point of Example 2.4.2 is the following: for linear models given in terms of transfer operators, as in (2.38), the optimal predictor is computed by inverting the noise model $H(q,\theta)$. By doing so, it is possible to reconstruct the process $\boldsymbol{e}_t$ if $\theta^\circ$ and the full history of the signals are known. Notice that the invertibility assumption is imposed as part of the model definition.

The computations of $\hat{\boldsymbol{y}}_{t|t-1}$ require the knowledge of the complete history of the input and output signals from $-\infty$ to $t-1$. In cases where the model is finite-dimensional (rational $G(q,\theta)$ and $H(q,\theta)$), it is sufficient to know the initial conditions exactly. If the initial conditions are unknown, they can be replaced by any reasonable guess (say 0). The resulting predictor will only be an approximation of the optimal predictor, but in most cases can be used to obtain acceptable solutions. The exact optimal predictor can be constructed by a time-varying linear filter computed using a recursive Kalman filter algorithm. For this, it is necessary to write the model in terms of a state-space representation and characterize the uncertainty regarding the initial conditions (history of the inputs and disturbances) using a probabilistic prior over the initial state.

**Example 2.4.3** (PEM through a time-varying Kalman filter)**.** Consider the following parameterized linear time-invariant state-space model

$$\begin{aligned} \boldsymbol{x}_{t+1} &= A(\theta)\boldsymbol{x}_t + B(\theta)u_t + \boldsymbol{w}_t, \quad & \boldsymbol{w}_t &\sim p(\boldsymbol{w}_t), \quad & \boldsymbol{x}_0 &\sim p(\boldsymbol{x}_0) \\ \boldsymbol{y}_t &= C(\theta)\boldsymbol{x}_t + \boldsymbol{v}_t, & \boldsymbol{v}_t &\sim p(\boldsymbol{v}_t), & t &= 1,\dots,N. \end{aligned} \tag{2.44}$$

We assume that the process disturbance $\boldsymbol{w}$ and the measurement noise $\boldsymbol{v}$ are both independent processes, mutually independent and independent of $\boldsymbol{x}_0$.

It is interesting to first observe that the model in its current form is not easy to invert. To clarify this point, let us assume that the initial state is known; that is $p(\boldsymbol{x}_0) = \delta_{x^0}$. The question is then how to use the data to reconstruct $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ or a function of them. We might first try to write the model in terms of transfer operators. For a given $\theta$,

$$\boldsymbol{y}_t = \underbrace{C(\theta)(qI - A(\theta))^{-1}B(\theta)u_t}_{\text{known function of } \theta} + \underbrace{C(\theta)(qI - A(\theta))^{-1}\boldsymbol{w}_t + \boldsymbol{v}_t}_{\text{stochastic with dynamics}}. \tag{2.45}$$

The difficulty seems to arise from the way the second term is written. Ideally, we would like to write the stochastic part in terms of an independent process or a martingale difference process which is unpredictable:

$$\boldsymbol{y}_t = \overbrace{\psi(\boldsymbol{D}_{t-1}, t,\ \theta)}^{\text{known function of } \theta \text{ at time } t} + \underbrace{\varepsilon_t}_{\text{stochastic and unpredictable}}.$$

Hence, the problem would be easier if it is possible to write the stochastic part involving the process disturbance and the measurement noise in terms of filtered versions of an independent process $\boldsymbol{\varepsilon}_t$. This requires a change of coordinates for the model, and it is well understood that this transformation is what the time-varying Kalman filter does (see [2, 53, 70]).

The required form is known as the innovations form and is given by the equations

$$\begin{aligned}\hat{\boldsymbol{x}}_{t+1}(\theta) &= A(\theta)\hat{\boldsymbol{x}}_t(\theta) + B(\theta)u_t + K_t(\theta)\boldsymbol{\varepsilon}_t\\ \boldsymbol{y}_t &= C(\theta)\hat{\boldsymbol{x}}_t(\theta) + \boldsymbol{\varepsilon}_t,\\ &\boldsymbol{\varepsilon}_t \text{ is a zero mean independent process with } \mathbf{cov}(\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_t) = \Lambda_t(\theta)\end{aligned} \tag{2.46}$$

in which the matrix $K_t(\theta)$ is known as the Kalman gain, and $\boldsymbol{\varepsilon}_t$ is the innovation in $\boldsymbol{y}_t$. The process $\boldsymbol{\varepsilon}$ is in fact the innovation process, see Definition 2.1.5. It is now easy to define the (time-varying) one-step ahead predictor, because the model as written in (2.46) is easy to invert. The optimal predictor is

$$\begin{aligned}\hat{\boldsymbol{y}}_{t|t-1}(\theta) &= C(\theta)\hat{\boldsymbol{x}}_t(\theta), \quad t = 1, \ldots, N,\\ \hat{\boldsymbol{x}}_t(\theta) &= A(\theta)\hat{\boldsymbol{x}}_{t-1}(\theta) + B(\theta)u_{t-1} + K_{t-1}(\theta)(\boldsymbol{y}_{t-1} - C(\theta)\hat{\boldsymbol{x}}_{t-1}(\theta)).\end{aligned} \tag{2.47}$$

When both the process noise and the measurement noise are Gaussian processes and the initial state is a Gaussian random vector, the innovation process $\boldsymbol{\varepsilon}$ is an independent Gaussian process with covariances $\Lambda_t(\theta)$. In this case, the PEM problem based on the time-varying Kalman filter and the norm

$$\ell(\boldsymbol{\varepsilon}_t(\theta), t; \theta) = \boldsymbol{\varepsilon}_t^{\top}(\theta)\Lambda_t^{-1}(\theta)\boldsymbol{\varepsilon}_t(\theta) + \log\det(\Lambda_t(\theta)) \tag{2.48}$$

coincides with the ML estimator. Observe that $\ell$ here is time-dependent and is parameterized by $\theta$ through the covariance matrices. Another PEM instance is obtained by using the optimal predictor and the Euclidean norm

$$\ell(\boldsymbol{\varepsilon}_t(\theta), t; \theta) = \|\boldsymbol{\varepsilon}_t(\theta)\|_2^2 = \boldsymbol{\varepsilon}_t^\top(\theta)\boldsymbol{\varepsilon}_t(\theta) \tag{2.49}$$

which is both time- and parameter-independent.

As we discussed above, choosing a different predictor and/or a different $\ell$ will result in different instances of the PEM family. Some of these choices might lead to consistent but suboptimal estimators, while others will not be consistent. It is a well-known fact, as shown in [92] for example, that for linear systems operating in open-loop and when $G$ and $H$ are parameterized independently, a predictor based on a misspecified $H$ will not affect the consistency of the PEM estimator of $G$. This is clarified in the following example.

**Example 2.4.4** (Ignoring process noise for LTI models of systems operating in open-loop)**.** Assume that the data is generated according to the model structure

$$\boldsymbol{y}_t = G(q, \theta^\circ)\boldsymbol{u}_t + \boldsymbol{\zeta}_t \tag{2.50}$$

in which $\boldsymbol{u}$ and $\boldsymbol{\zeta}$ are independent stationary processes with spectra $\Phi_u(w)$ and $\Phi_\zeta(w)$ respectively. Assume that $\boldsymbol{\zeta}$ is colored (this could be for example the case of a linear state-space model with process noise as in (2.45)). To construct a PEM estimator, we need to define a predictor and a norm. Consider the OE suboptimal predictor defined in (2.42) which ignores any predictable features of the process $\boldsymbol{\zeta}$, and let us pick $\ell(\cdot) = \|\cdot\|_2^2$ which is both time- and parameter-independent. Then, the PEM problem is defined by the unweighted nonlinear least-squares problem

$$\begin{aligned} \min_{\theta\in\Theta} \quad & \sum_{t=1}^{N} \|\boldsymbol{e}_t(\theta)\|_2^2 \\ \text{such that} \quad & \boldsymbol{e}_t(\theta) = \boldsymbol{y}_t - \hat{y}_{t|t-1}(\theta), \\ & \hat{y}_{t|t-1}(\theta) = G(q,\theta)u_t, \quad t = 1, \dots, N. \end{aligned} \tag{2.51}$$

The limiting estimate of the resulting estimator can be checked by applying Parseval's relation to the cost function (see [92, Section 8.5, page 266]); it holds that

$$\theta^* = \arg\min_{\theta\in\Theta} \int_{-\pi}^{\pi} |G(e^{i\omega}, \theta^\circ) - G(e^{i\omega}, \theta)|^2 \Phi_u(\omega)\,\mathrm{d}\omega + \int_{-\pi}^{\pi} \Phi_\zeta(\omega)\,\mathrm{d}\omega.$$

It is evident that when $\Phi_u(\omega) > 0 \ \forall\omega$, and under the identifiability condition

$G(e^{i\omega}, \theta^\circ) - G(e^{i\omega}, \theta) = 0$ for almost all $\omega$ (with respect to $\Phi_u(\omega)\,\mathrm{d}\omega$) $\Rightarrow \ \theta^\circ = \theta$

it holds that $\theta^* = \theta^\circ$. Therefore, the estimator defined by (2.51) for data coming from (2.50) is consistent.

### PEM when it is not ML

Even though the estimator defined by the PEM in Example 2.4.4 is consistent, it does not coincide with the MLE. We note here that the predictor is defined using only the first moment of $\boldsymbol{y}$ and an exact full probabilistic model is not required. It is of interest to observe that the PEM still solves a likelihood problem which happens to be misspecified. To clarify this point, we look at the choices made by the PEM (in terms of the predictor and the used norm) and examine their implications if the PEM problem is to coincide with an ML problem.

First, for the used predictor to coincide with a conditional mean, the data should be generated according to a data-generation mechanism

$$\boldsymbol{y}_t = G(q,\theta)u_t + \tilde{\boldsymbol{\zeta}}_t$$

in which $\tilde{\boldsymbol{\zeta}}$ is a zero mean independent process. Second, for the objective function of the PEM problem (2.51) to match a likelihood function, the relation (2.37) has to be satisfied. This occurs if $\tilde{\boldsymbol{\zeta}}$ is a Gaussian process with a constant variance which is $\theta$-independent; that is (for a scalar signal)

$$p(\tilde{\boldsymbol{\zeta}}_t) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\|\tilde{\boldsymbol{\zeta}}_t\|_2^2\right), \quad \forall t \in \mathbb{Z}. \tag{2.52}$$

This would imply that $\boldsymbol{y}$ is an independent Gaussian process with a mean function $m_t(\theta) = G(q,\theta)u_t$ and a constant ($\theta$-independent) variance. Under this model, the (misspecified) likelihood function of $\theta$ is

$$\tilde{p}(\boldsymbol{Y};\theta) = \prod_{t=1}^{N} p_{\tilde{\zeta}}(\boldsymbol{y}_t - G(q,\theta)u_t). \tag{2.53}$$

Solving the misspecified ML estimation problem is equivalent to solving the minimization problem

$$\min_{\theta} \sum_{t=1}^{N} \|\boldsymbol{y}_t - G(q,\theta)u_t\|_2^2 \tag{2.54}$$

where all the $\theta$-independent terms are dropped. This problem coincides with problem (2.51) formulated by the PEM. In Chapter 4, we will extend this notion to stochastic nonlinear models.

Before we summarize this chapter, we formulate the main problem of the thesis.

## 2.5 Problem Formulation

Assume that a finite data set $\boldsymbol{D}_N$ of known inputs and observed outputs, as defined in (2.29), is given such that

1. the outputs are generated according to a discrete-time stochastic parametric nonlinear dynamical model, as defined in Definition 2.10, with a (true) parameter $\theta^\circ \in \Theta$.
2. the unobserved stochastic process $\{\boldsymbol{\zeta}_k\}$ follows a well-defined distribution which can be parameterized (possibly independently) by $\theta^\circ \in \Theta$.

The objective is to construct (approximations of) consistent estimators of $\theta^\circ$,

$$\boldsymbol{D}_N \mapsto \hat{\theta}(\boldsymbol{D}_N).$$

The focus lies on the Maximum Likelihood Estimator (Definition 2.4.2), and on possible consistent instances of the PEMs family (Definition 2.4.4).

## 2.6 Summary

In this chapter, we summarized the necessary background and made several important remarks. In Section 2.1, we introduced a stochastic framework for the signals and defined both parametric linear and nonlinear models in discrete-time. We presented Wold's decomposition of a class of non-stationary processes in Theorem 2.1.6. In Section 2.3, we defined estimators and their properties. In Section 2.4, we defined the two estimation methods that concern us: the Maximum Likelihood Method and the family of Prediction Error Methods. We presented several examples of both methods for parametric linear models. Finally, in Section 2.5, we formulated the main problem of the thesis.

Chapter 3

# Approximate Solutions to the ML Problem

As argued in Chapter 1, the parameter identification problem of the general models defined in (2.10) is challenging due to the intractability of the objective functions defining the estimators. In this chapter, we will explore approaches to approximate solutions to the MLE problem, or equivalently the optimal instance of the PEMs. We first start in Section 3.2 by presenting the methods that can be used for tractable models. The goal is to identify the intractable quantities that would require approximations. In the latter sections, we study several possible analytic and numerical approximate algorithms. The intention here is not to promote the use of these methods, but to understand the fundamental difficulties of the problem.

## 3.1 Introduction

Recall that the MLE and the PEM estimators are defined by an optimization problem,

$$\hat{\boldsymbol{\theta}} := \arg\max_{\theta \in \Theta} \mathcal{J}(\boldsymbol{D}_N, \theta)$$

for some objective (cost) function $\mathcal{J}$ (in the case of the PEMs, it is equivalent to maximize the negative of the sum in (2.33)). Our goal is to study approximate solutions to the problem when the objective function is analytically intractable due to the involved multidimensional integrals (see Section 1.2.2).

It is important to understand that we are not interested in the objective function itself, but in the set of (global) maximizers. Note that even for tractable models, where $\mathcal{J}$ can be written in closed-form, the maximizers are usually not available in closed-form due to complex model parameterizations. Hence, numerical optimization routines are required in general. Methods applicable to intractable models combine iterative numerical optimization with analytic or numerical approximation of intractable quantities.

In this thesis, we are only considering a subclass of numerical optimization algorithms based on local explorations, known as hill-climbing methods. These are iterative numerical optimization methods that start at an initial feasible point

$\theta^{(0)} \in \Theta$, then attempt to find (according to a well-defined mathematical strategy) another candidate point in the parameter space. The method iterates over this parameter update step until convergence. In practice, convergence means that no further improvement can be achieved, a specific tolerance for some condition is met or a maximum number of iterations is reached. Such methods do not guarantee that the algorithm returns a global maximizer; they can only guarantee convergence to a local maximizer, or sometimes only to a stationary point. For this reason, the initial point $\theta^{(0)}$ has to be chosen carefully, and initialization methods with the sole purpose of finding a good $\theta^{(0)}$ are usually needed.

The mathematical formalism of iterative numerical optimization methods assigns to each algorithm a point-to-set mapping $\theta \mapsto \mathcal{A}(\theta) \subset \Theta$, see for example [94, Chapter 7]. The algorithm then generates a sequence of points $\{\theta^{(i)}\}$ according to the iterations $\theta^{(i+1)} \in \mathcal{A}(\theta^{(i)})$ in which $i \in \mathbb{N}_0$. From this point of view, it is clear that the key for achieving the aim that we set for ourselves (see Section 2.5) is the mapping $\mathcal{A}$ and not the objective function $\mathcal{J}$ itself. The hill-climbing methods considered in this thesis are algorithms that are applicable when the likelihood function and its gradient are available in closed-form. They can be divided into direct optimization methods or Minorization-Maximization methods.

Direct optimization methods can be divided into: (i) derivative-free algorithms, and (ii) gradient-based algorithms. In derivative-free optimization, the mapping $\mathcal{A}$ relies only on evaluations of the objective function at several points that depend on the current value $\theta^{(i)}$. In this case, approximating $\mathcal{A}$ requires the approximation of the objective function at the given points. A commonly used derivative-free algorithm is the Nelder-Mead algorithm (see [103]). On the other hand, the mapping $\mathcal{A}$ of gradient-based algorithms relies on the gradient of the objective function and possibly on the Hessian matrix. Approximating $\mathcal{A}$ in this case requires at least approximations of the gradient of the objective function. Depending on the algorithm, it might also need evaluations of the objective function itself. A commonly used gradient-based algorithm is the quasi-Newton algorithm (see [94, Chapter 10]).

An alternative to direct optimization methods is methods based on Minorization-Maximization algorithms, such as the celebrated Expectation-Maximization algorithm, originally introduced in [29]. These methods replace the original objective function by a surrogate function supposedly easier to maximize. The surrogate function is chosen such that the sequence $\{\theta^{(i)}\}$ converges to a local maximizer of the original objective function.

In any case, the approximate solutions are developed by approximating the mapping $\mathcal{A}$ of the used algorithm either analytically or numerically. They may require an approximation of the objective function, its gradient or another quantity like a surrogate function for example.

In this chapter, we are only considering the MLE problem. As shown in Chapter 2, the optimal instance of the PEMs is related to the MLE; hence, all the developed approximations in this chapter can also be seen as approximations of the optimal instance of the PEMs. Our goal is to study possible approximations of the MLE for intractable models. As shown in the next section, approximations of the MLE require approximations of PDFs over high-dimensional spaces, which makes the problem computationally expensive. Later in Chapter 4, we will describe a PEM that can be used to construct computationally attractive consistent estimators.

We will constrain the general model

$$\boldsymbol{y}_t = f_t(\{u_k\}_{k=1}^{t-1}, \{\boldsymbol{\zeta}_k\}_{k=1}^{t}; \theta), \quad t = 1, 2, \ldots, N,$$

given in Definition 2.1.8, to cases where the unobserved process is

$$\boldsymbol{\zeta}_t = \begin{bmatrix} \tilde{\boldsymbol{\zeta}}_t^\top & \boldsymbol{w}_t^\top & \boldsymbol{v}_t^\top \end{bmatrix}^\top,$$

in which $\boldsymbol{w} = \{\boldsymbol{w}_t\}$ and $\boldsymbol{v} = \{\boldsymbol{v}_t\}$ are independent and mutually independent white noise. The process $\boldsymbol{w}$ models any latent process (such as the process disturbance in state-space models), and the process $\boldsymbol{v}$ represents measurement noise. If required, $\tilde{\boldsymbol{\zeta}}$ can be used to model unknown initial conditions (like $\boldsymbol{x}_0$ for state-space models) and it is assumed to be independent of both $\boldsymbol{w}$ and $\boldsymbol{v}$.

To be able to simplify the exposition, we will assume that $\tilde{\boldsymbol{\zeta}}_t$ is known for all $t \in \mathbb{Z}$ or lumped together with $\boldsymbol{w}_t$. Define the column vectors

$$\boldsymbol{Y} := \begin{bmatrix} \boldsymbol{y}_1^\top & \ldots & \boldsymbol{y}_N^\top \end{bmatrix}^\top \in \mathsf{L}_2^{d_y N}, \qquad U := \begin{bmatrix} u_1^\top & \ldots u_{N-1}^\top \end{bmatrix}^\top \in \mathbb{R}^{d_u(N-1)},$$
$$\boldsymbol{W} := \begin{bmatrix} \boldsymbol{w}_1^\top & \ldots & \boldsymbol{w}_N^\top \end{bmatrix}^\top \in \mathsf{L}_2^{d_w N}, \qquad \boldsymbol{V} := \begin{bmatrix} \boldsymbol{v}_1^\top & \ldots \boldsymbol{v}_N^\top \end{bmatrix}^\top \in \mathsf{L}_2^{d_y N}.$$

and note that we will omit the dependence on the inputs $U$ from all the notations.

The methods of this chapter require the following assumption:

**Assumption 3.1.1.** *The joint PDF*

$$p(\boldsymbol{Y}, \boldsymbol{W}; \theta) \tag{3.1}$$

*has a known analytical form, parameterized by $\theta \in \Theta$.*

This assumption is not very restrictive, however. It holds for a wide range of interesting models; for example, state-space models with known forms for the state and the output equation and known PDFs for the initial state, process disturbance, and the measurement noise (see (2.12) and (2.17)). It may also be satisfied for non-Markovian models, such as switching systems and conditionally linear Gaussian state-space models (see [20]).

Observe that, in general, the joint PDF in (3.1) can be factored as

$$p(\boldsymbol{Y}, \boldsymbol{W}; \theta) = p(\boldsymbol{Y} | \boldsymbol{W}; \theta) p(\boldsymbol{W}; \theta). \tag{3.2}$$

The assumption that $\boldsymbol{\zeta}$ follows a known distribution (see Section 2.5) implies that $p(\boldsymbol{W}; \theta)$ is a known PDF, and Assumption 3.1.1 requires the conditional PDF $p(\boldsymbol{Y} | \boldsymbol{W}; \theta)$ to be known. This holds in the following scenarios:

1. The model is given by the relations

$$\boldsymbol{y}_t = f_t(\{u_k\}_{k=1}^{t-1}, \{\boldsymbol{w}_k\}_{k=1}^{t}; \theta) + \boldsymbol{v}_t, \quad t = 1, \dots, N,$$

   i.e., the measurement noise is additive. In this case, the output vector can be written in terms of a well-defined mapping $\mathcal{M}$,

$$\boldsymbol{Y} = \mathcal{M}(U, \boldsymbol{W}; \theta) + \boldsymbol{V},$$

   and for any realizations $Y$ and $W$

$$p(Y|W; \theta) = p_{\boldsymbol{V}}(Y - \mathcal{M}(U, W; \theta); \theta).$$

2. If the measurement noise $\boldsymbol{v}$ is not additive, we assume that it is possible to evaluate a derived (conditional) density for $\boldsymbol{Y}$ given $\boldsymbol{W}$; see for example [64, Theorem 2.7]. This assumption restricts the models to those that can be (conditionally) inverted with respect to $\boldsymbol{v}_t$.

Because $\boldsymbol{W}$ is not observed, the PDF of the output vector $\boldsymbol{Y}$ has to be calculated by marginalization, namely

$$p(\boldsymbol{Y}; \theta) = \int_{\mathbb{R}^{d_w N}} p(\boldsymbol{Y}, W; \theta) \, \mathrm{d}W. \tag{3.3}$$

Using (3.2), we have

$$p(\boldsymbol{Y}; \theta) = \mathbb{E}_{\boldsymbol{W}} \left[ p(\boldsymbol{Y}|\boldsymbol{W}; \theta); \theta \right],$$

in which we used the notation $\mathbb{E}_{\boldsymbol{W}}[\cdot; \theta]$ to denote the expectation with respect to the distribution of the random quantity $\boldsymbol{W}$. The notation indicates that, in general, the distribution of $\boldsymbol{W}$ is $\theta$-dependent. The ML estimate (see Definition 2.4.2), is then given by

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \, \mathbb{E} \left[ p(Y|\boldsymbol{W}; \theta); \theta \right]$$

where the function to be maximized is defined by a multidimensional integral and is generally not available in closed-form.

### A key quantity

A key quantity for most of the algorithms is the posterior PDF of $\boldsymbol{W}$ which we denote by $p(\boldsymbol{W}|Y; \theta)^1$. To appreciate the importance of this PDF, observe that

$$p(Y; \theta) = \frac{p(Y, \boldsymbol{W}; \theta)}{p(\boldsymbol{W}|Y; \theta)} \tag{3.4}$$

for any feasible $Y$ and $\theta$, and that the right-hand side is independent of any specific value of $\boldsymbol{W}$. Thus, if the PDF $p(\boldsymbol{W}|Y; \theta)$ is known, the likelihood of $\theta$ for any given realization $Y$ is given by the fraction $p(Y, W; \theta)/p(W|Y; \theta)$ with any arbitrary value $W$ in the support of $p(\boldsymbol{W}; \theta)$. Under Assumption 3.1.1, the numerator of this fraction is known. Consequently, determining the likelihood function is as hard as determining the posterior of $\boldsymbol{W}$. As we will shortly see, most of the algorithms shift the difficulty of the problem from the likelihood function to the posterior of $\boldsymbol{W}$.

Because neither $\theta$ nor $p(\boldsymbol{W}|Y;\theta)$ are known, most of the algorithms iterate conditionally between both: conditioned on a candidate value $\theta^{(i)}$ the algorithms compute (an approximation of) $p(\boldsymbol{W}|Y;\theta^{(i)})$ or a value for $\boldsymbol{W}$ and then use it to generate the next candidate $\theta^{(i+1)}$, and so on.

## 3.2 Algorithms for Tractable Models

Before discussing any approximation approaches, we describe two main algorithm types that can be used to solve the ML (or the Maximum A-Posteriori) problem when the posterior of $\boldsymbol{W}$, or the likelihood function and its gradient, possess a known analytic form: the Expectation-Maximization algorithms and the gradient-based algorithms. In some cases, one of the two types might be preferred over the other. For example, if the model is simple enough, the Expectation-Maximization algorithm can have closed-form iterations and does not require the explicit evaluation of the gradient vector; however, it is also known that it is much slower compared to the quasi-Newton algorithm for example.

### 3.2.1 The Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm is an iterative (hill-climbing) algorithm, used to solve the ML estimation problem when the likelihood function is written in terms of a marginalization integral (see for example (3.3)). The algorithm is originally due to [29], it was published in 1977 and has been extensively used in the statistical inference community (cited by more than 50,000 articles/books[2]).

The main idea of the EM algorithm is to break the main optimization problem into two simpler related problems. In the first one, it is assumed that $\theta$ is known and the algorithm computes a distribution for the unobserved (missing) vector $\boldsymbol{W}$, and in the second one, it is assumed that such a distribution is given and the algorithm computes a value for $\theta$ by solving a maximization problem. To explain the idea, first observe that

$$p(\boldsymbol{W}|Y;\theta) = \frac{p(Y,\boldsymbol{W};\theta)}{p(Y;\theta)}, \tag{3.5}$$

and therefore

$$\log p(Y;\theta) = \log p(Y,\boldsymbol{W};\theta) - \log p(\boldsymbol{W}|Y;\theta). \tag{3.6}$$

[1] For brevity, we will refer to this PDF as “the posterior of $\boldsymbol{W}$”. It is also known as “the smoothing density of $\boldsymbol{W}$”.

[2] Estimated by Google Scholar’s index as of November 2017.

Assume that a value $\theta^{(i)}$ is given and use the model to evaluate the conditional density $p(\boldsymbol{W}|Y;\theta^{(i)})$. By integrating (3.6) with respect to $p(\boldsymbol{W}|Y;\theta^{(i)})$ we get that

$$\log p(Y;\theta) = \underbrace{\mathbb{E}[\log p(Y,\boldsymbol{W};\theta)|Y;\theta^{(i)}]}_{=:Q(\theta,\theta^{(i)})} - \underbrace{\mathbb{E}[\log p(\boldsymbol{W}|Y;\theta)|Y;\theta^{(i)}]}_{=:\mathcal{V}(\theta,\theta^{(i)})}. \tag{3.7}$$

The quantity

$$Q(\theta,\theta^{(i)}) = \int_{\mathbb{R}^{d_w N}} \log p(Y,W;\theta)p(W|Y;\theta^{(i)})\,\mathrm{d}W \tag{3.8}$$

is known as the intermediate quantity (or the $Q$-function) of the EM algorithm. It is a real-valued function over $\Theta$, indexed by $\theta^{(i)}$. The step of evaluating the intermediate quantity is known as the Expectation step (E-step) of the algorithm. The quantity

$$-\mathcal{V}(\theta,\theta^{(i)}) = \int_{\mathbb{R}^{d_w N}} -\log p(W|Y;\theta)p(W|Y;\theta^{(i)})\,\mathrm{d}W$$

is known as the entropy of $p(\boldsymbol{W}|\boldsymbol{Y};\theta^{(i)})$. The increment $\big(\mathcal{V}(\theta^{(i)},\theta^{(i)}) - \mathcal{V}(\theta,\theta^{(i)})\big)$ is the Kullback-Leibler divergence (relative entropy) between $p(\boldsymbol{W}|\boldsymbol{Y};\theta^{(i)})$ and $p(\boldsymbol{W}|\boldsymbol{Y};\theta)$ and is always nonnegative, see [24] for example.

Using (3.7), it holds that

$$\log p(\boldsymbol{Y};\theta) - \log p(\boldsymbol{Y};\theta^{(i)}) = \big(Q(\theta,\theta^{(i)}) - Q(\theta^{(i)},\theta^{(i)})\big) + \big(\mathcal{V}(\theta^{(i)},\theta^{(i)}) - \mathcal{V}(\theta,\theta^{(i)})\big).$$

Therefore, for every value $\theta^{(i+1)}$ such that $Q(\theta^{(i+1)},\theta^{(i)}) - Q(\theta^{(i)},\theta^{(i)}) \geq 0$ it holds that

$$\log p(Y;\theta^{(i+1)}) - \log p(Y;\theta^{(i)}) \geq 0. \tag{3.9}$$

The EM algorithm as introduced in [29] defines $\theta^{(i+1)}$ as the global maximizer of $Q(\theta,\theta^{(i)})$ over $\Theta$. This guarantees that the inequality (3.9) is satisfied and it means that the EM algorithm is a monotone optimization algorithm. The step of computing $\theta^{(i+1)}$ is known as the Maximization step (M-step) of the algorithm. Iterating over the above two steps will result in a sequence $\{\theta^{(i)}\}$ mapped by the likelihood function into a nondecreasing sequence of positive reals.

In summary, the EM algorithm is given by the iterations $\theta^{(i+1)} \in \mathcal{A}(\theta^{(i)})$ where

$$\mathcal{A}(\theta^{(i)}) := \{\theta \in \Theta : \theta \in \arg\max_{\theta\in\Theta} Q(\theta,\theta^{(i)})\}. \tag{3.10}$$

The procedure is illustrated in Figure 3.1 and summarized in Algorithm 1. We summarize the monotonicity property of the algorithm in the following theorem.

**Theorem 3.2.1** (Monotonicity of the EM algorithm)**.** *The sequence $\{\theta^{(i)}\}$ defined by Algorithm 1 satisfies*

$$p(Y;\theta^{(i+1)}) \geq p(Y;\theta^{(i)}), \text{ for every realization } Y, \ \ i \in \mathbb{N}_0$$

*and the equality holds if and only if $Q(\theta^{(i+1)},\theta^{(i)}) = Q(\theta^{(i)},\theta^{(i)})$.*

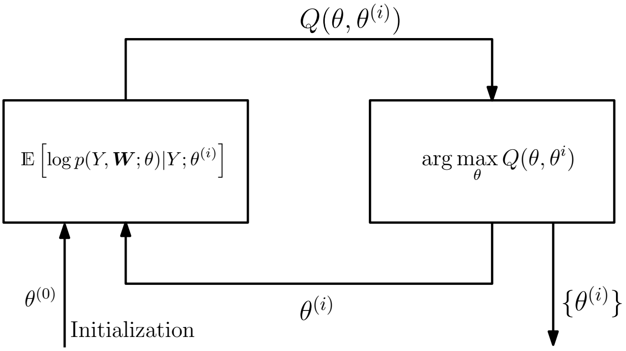

$Q(\theta, \theta^{(i)})$
$\mathbb{E}\left[\log p(Y, \boldsymbol{W}; \theta) | Y; \theta^{(i)}\right]$
$\arg\max_{\theta} Q(\theta, \theta^{i})$
$\theta^{(0)}$
Initialization
$\theta^{(i)}$
$\{\theta^{(i)}\}$

**Figure 3.1:** Illustration of the EM algorithm.

*Proof.* The proof (sketched above) is due to Dempster et al. [29]. Also see [98]. ■

Notice that the theorem does not guarantee that the sequence $\{\theta^{(i)},\ i \in \mathbb{N}_0\}$ converges to a maximum likelihood estimate. Further conditions on the mapping $\mathcal{A}$ in (3.10) are required to ensure the convergence to a stationary point. For details on these conditions we refer the reader to [16, 149]. Observe however that even under such conditions, the convergence is guaranteed only to a stationary point of the likelihood function. Additional conditions on the intermediate quantity can guarantee the convergence to a local maximum, but the available conditions are not easy to verify in practice. In any case, the EM algorithm is usually run several times with different (may be arbitrary) starting points $\theta^{(0)}$ to avoid convergence to undesirable stationary points.

**Algorithm 1:** The Expectation-Maximization (EM) algorithm [29].

```
input  : an initial guess $\theta^{(0)}$, the data $(Y, U)$, the conditional (smoothing)
         PDF $p(\boldsymbol{W}|Y;\theta)$, and a convergence (stopping) criterion
output : an approximate local maximum of the likelihood function $\hat{\theta}$
Initialize $\theta^{(i)}$ and set index $i \leftarrow 0$
while not converged do
    E-step: Compute $Q(\theta, \theta^{(i)}) = \mathbb{E}[\log p(Y, \boldsymbol{W}; \theta) | Y; \theta^{(i)}]$
    M-step: Compute $\theta^{(i+1)} = \arg\max_{\theta \in \Theta} Q(\theta, \theta^{(i)})$
    $i \leftarrow i + 1$
end
Set $\hat{\theta} = \theta^{(i)}$
```

### 3.2.2 Gradient-based Algorithms

An alternative to the EM algorithm is the family of gradient-based optimization methods. Two of the many comprehensive references on general numerical optimization algorithms are [108] and [94].

In this part, we will briefly discuss (i) the steepest ascent algorithm, and (ii) Newton's algorithm. Both are iterative hill-climbing algorithms that target local solutions. The focus is on describing the quantities needed within the iterations of the algorithms. We will only look at unconstrained problems and assume the existence of the gradient vector $\nabla_\theta p(Y;\theta)$ and the Hessian matrix $\nabla^2_\theta p(Y;\theta)$ of the likelihood function with respect to the vector $\theta$. To keep the notation uncluttered, we will use the notation $\nabla_\theta p(Y;\theta^{(i)})$ to denote the value of the gradient at $\theta^{(i)}$ instead of the explicit notation $\nabla_\theta p(Y;\theta)|_{\theta=\theta^{(i)}}$. A similar notation is used for the Hessian matrix.

#### The steepest ascent algorithm

The steepest ascent algorithm is one of the simplest gradient-based methods for optimization. It is based on the recursive formula

$$\theta^{(i+1)} = \theta^{(i)} + \alpha\nabla_\theta p(Y;\theta^{(i)})$$

in which the step size $\alpha$ is a small non-negative real number. For such a fixed step size, the algorithm only needs evaluations of the gradient. However, to guarantee the global convergence, i.e., convergence for all possible starting points $\theta^{(0)}$, with optimal rate, the step size $\alpha$ should be adaptively computed for each iteration. Ideally, it should be computed by solving the line-search problem

$$\begin{aligned} \alpha_i &= \arg\max_{\alpha\in\mathbb{R}_+} \quad p(Y;\theta(\alpha)) \\ &\quad \text{such that} \quad \theta(\alpha) = \theta^{(i)} + \alpha\nabla_\theta p(Y;\theta^{(i)}). \end{aligned} \tag{3.11}$$

However, solving this optimization problem for each iteration is computationally expensive. Notice that, in general, a solution to this problem will not be available in closed-form. Furthermore, it requires the evaluation of the likelihood function at candidate values $\theta(\alpha)$.

In practice, an inexact line-search method, like backtracking (see [108, Chapter 3]), is used in almost every case. Inexact algorithms check a given set of conditions at several values of $\alpha$ and select the first value that satisfies the conditions. For this, evaluations of the likelihood function and its gradient are required.

Another possible method for step size selection is the Barzilai and Borwein method [9] which solves, instead of (3.11), a quadratic problem (in $\alpha$) with a closed-form solution. Even though this leads to a non-monotone method, global convergence can be established in the quadratic case (in $\theta$, see [9]). In the general case however, the step size computed using Barzilai and Borwein method may be unacceptable and it must be modified to establish the convergence of the algorithm (see [56]). For this, evaluations of the likelihood function and its gradient are again required.

In summary, the algorithm is given by the iterations $\theta^{(i+1)} \in \mathcal{A}(\theta^{(i)})$ where

$$\mathcal{A}(\theta^{(i)}) := \{\theta \in \Theta : \theta = \theta^{(i)} + \alpha_i \nabla_\theta p(Y;\theta^{(i)})\},$$

in which $\alpha_i$ is computed by an inexact line-search method. This may require many evaluations of the likelihood function for each iteration. The convergence rate of the steepest ascent is known to be slow (linear) and with access to evaluations of the likelihood function and its gradient, better and faster algorithms can be used.

### Newton's algorithm

Newton's method is based on a second-order approximation of the likelihood gradient;

$$\nabla_\theta p(Y;\theta) \approx \nabla_\theta p(Y;\tilde{\theta}) + \nabla^2_\theta p(Y;\tilde{\theta})(\theta - \tilde{\theta}). \tag{3.12}$$

The first order condition $\nabla_\theta p(Y;\theta) = 0$ is approximated by the relation $\theta = \tilde{\theta} - \nabla^2_\theta p^{-1}(Y;\tilde{\theta})\nabla_\theta p(Y;\tilde{\theta})$ and therefore the algorithm is based on the recursive formula

$$\theta^{(i+1)} = \theta^{(i)} - \left[\nabla^2_\theta p(Y;\theta^{(i)})\right]^{-1} \nabla_\theta p(Y;\theta^{(i)}).$$

To guarantee desirable convergence properties, see [108, Theorem 3.5], the method is usually modified by introducing a small step size $\alpha_i$ such that

$$\theta^{(i+1)} = \theta^{(i)} - \alpha_i \left[\nabla^2_\theta p(Y;\theta^{(i)})\right]^{-1} \nabla_\theta p(Y;\theta^{(i)}).$$

Similarly to the steepest ascent algorithm, the step size is to be computed for each iteration and it requires evaluations of the likelihood function at several candidate points $\theta(\alpha)$.

The convergence rate of Newton's algorithm is quadratic[3] and therefore is recognized to be very efficient. However, its application is hindered by a numerical difficulty.

The implementation of the Newton's algorithm is prone to numerical instability due to possible non-invertible or poorly conditioned Hessian matrices. In addition, computing and inverting the Hessian matrix adds to the required computational effort. Fortunately, there are several methods that can be used to alleviate such difficulties, one of which is the quasi-Newton method. The idea is to approximate the inverse of the Hessian matrix directly by a positive definite weight matrix $H_i$ computed based on gradient evaluations. One commonly used approximation is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) approximation, see [94, Chapter 10] or [108, Chapter 6]. Although such an approximation is based only on gradient evaluations, the line-search of the quasi-Newton algorithm requires the evaluation of the likelihood function at several candidate points.

[3] See [108, Section 3.3] for the definition and details of convergence rates of iterative optimization algorithms.

In summary, the algorithm is given by the iterations $\theta^{(i+1)} \in \mathcal{A}(\theta^{(i)})$ where

$$\mathcal{A}(\theta^{(i)}) := \{\theta \in \Theta : \theta = \theta^{(i)} + \alpha_i H_i \nabla_\theta p(Y;\theta^{(i)})\}$$

in which $\alpha_i$ is computed by an inexact line-search method, and $H_i$ is an approximation of the Hessian inverse.

The quasi-Newton algorithm is preferred over the steepest ascent algorithm for nonlinear problems due to its faster convergence. The algorithm is illustrated in Figure 3.2.

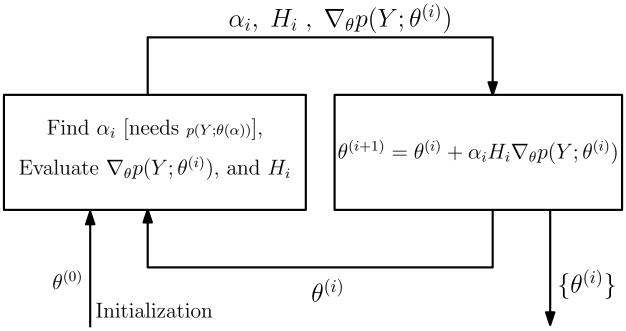

$\alpha_i,\ H_i\ ,\ \nabla_\theta p(Y;\theta^{(i)})$
Find $\alpha_i$ [needs $p(Y;\theta(\alpha))$],
Evaluate $\nabla_\theta p(Y;\theta^{(i)})$, and $H_i$
$\theta^{(i+1)} = \theta^{(i)} + \alpha_i H_i \nabla_\theta p(Y;\theta^{(i)})$
$\theta^{(0)}$
Initialization
$\theta^{(i)}$
$\{\theta^{(i)}\}$

**Figure 3.2:** Illustration of the quasi-Newton algorithm.

### Log-likelihoods versus Likelihoods

When gradient-based algorithms are used to solve the ML problem, it is usually recommended from a numerical point of view to work with the logarithm of the likelihood function (log-likelihood). We give two reasons for this.

The first is related to the tractability of the values of the objective function. To clarify this point, assume that the number of samples $N$ is large, and the output $\boldsymbol{y}_t$ is independent over time. Then for any realization $Y$ the value of the likelihood function at $\theta$ is given by $p(Y;\theta) = \prod_{t=1}^{N} p(y_t;\theta)$ in which, for several $t$, the value $p(y_t;\theta)$ might be small. This means that the value $p(Y;\theta)$ will be very small. If $N = 100$ and each term of the product is around 0.4, this value is $0.4^{100} \approx 1.6 \times 10^{-40}$ which is less than machine precision. The use of the logarithm turns the product into sums and produce tractable numbers; $100\log(0.4) \approx -91.62$.

The second reason is related to the scale of the gradient. The gradient of the log-likelihood function is usually well-scaled compared to the gradient of the likelihood function. This is evident in cases where $p(Y;\theta) \propto \exp(-\varphi(Y,\theta))$ for some positive function $\varphi$. In such cases, $\nabla_\theta p(Y;\theta) = -p(Y;\theta)\nabla_\theta \varphi(Y;\theta)$ will take very small values. On the other hand

$$\nabla_\theta \log p(Y;\theta) = \nabla_\theta p(Y;\theta)\frac{1}{p(Y;\theta)}, \tag{3.13}$$

which is a scaled version of the gradient of the likelihood function, and will usually assume tractable values.

As will become clear in what follows, regardless of the used algorithm, working with the log-likelihood function and its gradient requires the knowledge of the posterior PDF $p(\boldsymbol{W}|Y;\theta)$. Whenever the involved integrals are written with respect to the (known) prior $p(\boldsymbol{W};\theta)$, the involved quantities have to be multiplied by the conditional likelihood function $p(Y|W;\theta)$ which leads to numerical difficulties.

To conclude this part, we compare the EM algorithm to the quasi-Newton algorithm.

### The EM algorithm versus the quasi-Newton algorithm

As described above, the EM algorithm indicates that $p(\boldsymbol{W}|Y;\theta)$ is a key ingredient. Unfortunately, according to the assumed model, this PDF is not always available in closed-form. The EM algorithm is expected to be very helpful in cases where the following two conditions hold: (i) it is possible to evaluate the intermediate quantity (i.e., compute the E-step) at a reasonable computational cost, and (ii) the intermediate quantity has a sufficiently simple form to allow for a closed-form solution of the M-step. Usually, see [98, Section 1.5.3], these two conditions hold when the joint PDF $p(\boldsymbol{Y},\boldsymbol{W};\theta)$ is a member of the exponential family (see [123] for detailed definition and properties of the exponential family of distributions). In such cases, the intermediate quantity takes the form

$$Q(\theta,\theta^{(i)}) \propto \beta^{\top}(\theta)\mathbb{E}[s(Y,\boldsymbol{W})|Y;\theta^{(i)}] - c(\theta) \tag{3.14}$$

in which $s$ is a natural sufficient statistic, the E-step reduces to the conditional expectation of $s$, and the M-step is available in closed-form whenever

$$\arg\max_{\theta\in\Theta} \ \beta^{\top}(\theta)S - c(\theta)$$

takes a closed-form for every given vector $S$. In this situation, the EM algorithm is known to be parameterization-independent or scale-free. This means that for any one-to-one transformation of $\theta$, the EM iterations remain unchanged and the convergence rate of the EM algorithm is not affected. Furthermore, for cases with a simple M-step, it is possible to consider constraints implicitly by introducing Lagrange multipliers (see [94, Chapter 11]).

In cases where the E-step is computationally expensive and/or the M-step is a complicated maximization problem by itself, it is not clear whether the EM algorithm is advantageous or not. Nevertheless, observe that the EM algorithm does not really need a (local or global) maximizer in the M-step. Instead, it requires any point $\theta^{(i+1)}$ such that $Q(\theta^{(i+1)},\theta^{(i)}) > Q(\theta^{(i)},\theta^{(i)})$, see (3.9). Several options have been suggested in the literature to overcome the intractability of the M-step. One possibility is to use conditional maximization (see [99]). Another option is to use only one step of a Newton's method in the M-step (see [76]). It has also been noticed that the EM algorithm seem to avoid local minima of (erratic) log-likelihood functions (see [107, Section 3.4]). The argument used to explain this behavior is that the intermediate quantity forms a global approximation of the log-likelihood function, unlike the local approximations (see (3.12)) that are implicit to gradient-based algorithms.

In several other cases, the quasi-Newton algorithm might be preferred, especially if it does not require an additional maximization step within its iterations. Due to the use of (an easy to compute) approximation of the Hessian matrix, the quasi-Newton algorithm might also reach quadratic convergence and is usually faster than the EM algorithm when the dimension of $\theta$ is large or the variance of $\boldsymbol{W}$ is small. On the other hand, there exist several suggestions for accelerating the convergence of the EM algorithm using the Hessian matrix (see [76, 77] for example).

**Table 3.1:** Summary of the required quantities at each iteration for the Expectation-Maximization algorithm and the quasi-Newton algorithm.

| | Expectation-Maximization | quasi-Newton (BFGS) |
|---|---|---|
| Needed quantities per iteration | $p(\boldsymbol{W}\|\boldsymbol{Y};\theta^{(i)})$, $\arg\max_{\theta\in\Theta} Q(\theta,\theta^{(i)})$ (see (3.5) and (3.8)) | $\nabla_\theta p(Y;\theta^{(i)})$, $p(Y;\tilde{\theta})$ for several $\tilde{\theta}$ (see (3.3)) |

We have summarized the findings regarding the required quantities for the EM and quasi-Newton algorithms in Table 3.1. In what follows, we will discuss possible approximations of these quantities and test the corresponding algorithms on relatively simple models with analytically intractable likelihood functions. We first start with analytical approximations and then consider numerical approximations.

## 3.3 Analytical Approximations

All the values in Table 3.1 are given in terms of intractable multidimensional integrals. Analytical approximations of multidimensional integrals have long been used to approximate various quantities in mathematical physics and statistics. Once an approximation is obtained, the original problem usually simplifies considerably, and closed-form solutions based on the obtained approximation might be available. However, there is in general no way to exactly evaluate the accuracy of the resulting approximations.

Laplace's approximation method (named after Pierre-Simon Laplace (1749–1827)) is a technique that originally appeared in [78] where it aimed at approximating a particular instance of the one-dimensional integrals of the form

$$\int_a^b h(t)\exp\left(-\lambda M(t)\right)\mathrm{d}t, \quad a,b\in\mathbb{R}\cup\{\infty,-\infty\} \tag{3.15}$$

when $\lambda > 0$ is large. Since then, the method has been extended to similar multidimensional integrals in addition to contour integrations (see [37]). The basic idea of the approximation is that large contributions to the value of the integral occur at values of $t$ around the (assumed unique) minimizer of $M(t)$. Let us denote the minimizer by $t^\star$, then Laplace's approximation of (3.15), under some conditions (see [37, Section 4.3.3]), is given by the value

$$\sqrt{\frac{2\pi}{\lambda M''(t^\star)}}\, h(t^\star) \exp\left(-\lambda M(t^\star)\right).$$

In statistical inference, Laplace's approximation method has been used in a Bayesian framework to approximate posterior PDFs of parameter vectors of fixed dimension by using a multivariate Gaussian density (see [141] for example). This can be motivated by the asymptotic normality of posteriors that holds under some regularity conditions (see [123, Theorem 7.89]). Laplace's approximation method can also be used to approximate Bayes factors, which are used for model selection and comparison, by approximating the marginal PDF of the data (see [123, Theorem 7.166]) and may be used to motivate the Bayesian information criterion (see [15, Section 4.41]).

In this thesis, the ideas of Laplace's approximation are used to approximate both the posterior PDF $p(\boldsymbol{W}|Y;\theta)$ and the value of the likelihood function $p(Y,\theta)$ for a given $\theta$. Before describing the method, we make the following assumption.

**Assumption 3.3.1.** *For every given $Y$ and $\theta \in \Theta$, the function, $p(Y,W;\theta)$ is twice continuously differentiable with respect to $W$.*

Now, recall that Bayes' theorem states that

$$p(\boldsymbol{W}|Y;\theta) = \frac{p(Y,\boldsymbol{W};\theta)}{p(Y;\theta)}. \tag{3.16}$$

Under Assumption 3.1.1, the numerator in (3.16) is known in closed-form, but not the normalizing factor (the likelihood function).

To define an approximation of the posterior we proceed as follows. For a given candidate value $\theta$ and a realization $Y$, we first find a mode of the posterior by computing

$$\arg\max_W\; p(W|Y;\theta) = \arg\max_W\; \frac{p(Y,W;\theta)}{p(Y;\theta)} = \arg\max_W\; p(Y,W;\theta). \tag{3.17}$$

Let us denote any local maximizer[4] of (3.17) by $\widehat{W}(\theta)$ and any global maximizer by $\widehat{W}_{\mathrm{MAP}}(\theta)$. By definition, the vector $\widehat{W}_{\mathrm{MAP}}$ is a Maximum a Posteriori (MAP) estimate of the unobserved $\boldsymbol{W}$ (see [118, Section 4.1.2]). Note that the maximizer depends on $\theta$ (and of course on the realization $Y$). Furthermore, define the matrix

$$\Pi(\widehat{W}(\theta),\theta) \coloneqq \nabla_W^2 \log p(Y,W;\theta)\big|_{W=\widehat{W}(\theta)},$$

and observe that for every local maximizer $\widehat{W}(\theta)$ it must hold that $\Pi(\widehat{W}(\theta),\theta) \prec 0$.

[4] Also known as a "mode" of the PDF. A PDF is called "uni-modal" if it has a single mode. If it has two or more modes, it is called "bi-modal" or "multi-modal" respectively.

### The posterior PDF

The next step is to write the second-order Taylor approximation of $\log p(Y, W; \theta)$ around a local maximum $\widehat{W}(\theta)$. It holds that

$$\log p(Y, \boldsymbol{W}; \theta) \approx \log p(Y, \widehat{W}(\theta); \theta) + \frac{1}{2}(\boldsymbol{W} - \widehat{W}(\theta))^{\top}\Pi(\widehat{W}(\theta), \theta)(\boldsymbol{W} - \widehat{W}(\theta)).$$

Notice that the first-order term is equal to zero because $\widehat{W}(\theta)$ is a local maximum. This implies that

$$p(Y, \boldsymbol{W}; \theta) \approx p(Y, \widehat{W}(\theta); \theta) \exp\left(\frac{1}{2}(\boldsymbol{W} - \widehat{W}(\theta))^{\top}\Pi(\widehat{W}(\theta), \theta)(\boldsymbol{W} - \widehat{W}(\theta))\right). \quad (3.18)$$

Because $p(\boldsymbol{W}|Y; \theta)$, as a PDF for $\boldsymbol{W}$, is proportional to $p(Y, \boldsymbol{W}; \theta)$ (see (3.16)), we get an approximation of the posterior of $\boldsymbol{W}$ if we normalize the right-hand side of (3.18). Since the argument of the exponential function on the right-hand side is quadratic in $\boldsymbol{W}$, we can use standard results for the normalization of multivariate Gaussian distributions (see Appendix C) to obtain the Gaussian approximation

$$\tilde{p}(\boldsymbol{W}|Y; \theta) = \mathcal{N}(\widehat{W}(\theta), \Sigma(\widehat{W}(\theta), \theta)) \quad (3.19)$$

in which the mean is given by $\widehat{W}(\theta)$, and the covariance matrix is

$$\Sigma(\widehat{W}(\theta), \theta) := -\Pi^{-1}(\widehat{W}(\theta), \theta) \succ 0. \quad (3.20)$$

### The likelihood function

The approximation in (3.18) can also be used to obtain an approximation of the likelihood function of $\theta$ at $Y$ (the normalization constant in (3.16)). Recall that the likelihood function is given by the marginalization integral (3.3), seen as a function of $\theta$ with fixed realization $Y$. To get the required approximation, we instead solve the tractable integral

$$\int p(Y, \widehat{W}(\theta); \theta) \exp\left(\frac{1}{2}(W - \widehat{W}(\theta))^{\top}\Pi(\widehat{W}(\theta), \theta)(W - \widehat{W}(\theta))\right) \mathrm{d}W.$$

By using standard results on Gaussian integrals and observing that the first factor of the integrand is independent of $W$, we get the approximation

$$\tilde{p}(Y; \theta) = p(Y, \widehat{W}(\theta); \theta) \cdot \sqrt{(2\pi)^{d_w N}} \sqrt{\det\left(\Sigma(\widehat{W}(\theta), \theta)\right)}, \quad (3.21)$$

or alternatively

$$\log \tilde{p}(Y; \theta) = \log p(Y, \widehat{W}(\theta); \theta) + \frac{d_w N}{2} \log(2\pi) + \frac{1}{2} \log\det(\Sigma(\widehat{W}(\theta), \theta)). \quad (3.22)$$

The above approximations of the posterior of $\boldsymbol{W}$ and the likelihood function imply the following approximation for the joint PDF (evaluated at $Y$)

$$p(Y, \boldsymbol{W}; \theta) \approx \tilde{p}(Y, \boldsymbol{W}; \theta) := \tilde{p}(Y; \theta)\tilde{p}(\boldsymbol{W}|Y; \theta). \quad (3.23)$$

We summarize several observations on Laplace's approximation in the following remark.

**Remark 3.3.2** (Properties of Laplace's approximation)**.**

- *Laplace's approximations of $p(Y, \boldsymbol{W}; \theta)$ and $p(Y, \theta)$ are $\theta$-dependent. Therefore, when used in conjunction with an iterative optimization algorithm, such as the EM or the quasi-Newton algorithm, new approximations have to be computed for each new candidate $\theta^{(i)}$. To avoid evaluating (3.17) many times, algorithms that require the least number of iterations would be preferred.*

- *Laplace's approximation depends on how the unobserved process is modeled. For example, if the model has a latent (state) process $\boldsymbol{X}$ that depends on $\boldsymbol{W}$, the approximation of the likelihood function that we compute based on $p(Y, \boldsymbol{W}; \theta)$ (see (3.21)) is generally different from the one that would be obtained based on $p(Y, \boldsymbol{X}; \theta)$. Observe that if $\theta$ is considered known, one can look at $\boldsymbol{W}$ and $\boldsymbol{X}$ as different "parameterizations" of the model.*

- *Since the dimension of the vector $\boldsymbol{W}$ depends on the data size $N$, the posterior $p(\boldsymbol{W}|Y; \theta)$ does not concentrate as $N$ increases; it does not approach a Gaussian PDF for example, and according to the model it might well be a multi-modal PDF regardless of $N$.*

- *Laplace's approximation is exact if and only if $p(\boldsymbol{Y}, \boldsymbol{W}; \theta)$ is a multivariate Gaussian distribution (see Example 3.3.2).*

- *The basic idea behind Laplace's approximation assumes that the PDF to be approximated has one dominant mode. In case of multi-modal posterior PDFs, the approximation will only provide a local description around one mode.*

Next, we will investigate how the approximations (3.19) and (3.21) can be used to approximate quantities in Table 3.1. We start with the EM algorithm.

### 3.3.1 Approximate Expectation-Maximization

At iteration $i + 1$ of the EM algorithm, we are given $\theta^{(i)}$ and a corresponding approximation $\tilde{p}(\boldsymbol{W}|Y; \theta^{(i)})$. The E-step is the step of evaluating the integral defining the intermediate quantity, see (3.8). We will present two ways to proceed with an approximate solution; they rely on different approximations of the integrand of (3.8). To simplify the expressions of the intermediate quantity, we will make the following assumption.

**Assumption 3.3.3.** *The conditional PDF $p(\boldsymbol{Y}|W; \theta)$ is Gaussian with a mean vector $\mu(W; \theta)$ and a covariance matrix $\Sigma_Y(\theta)$. Furthermore, the vector $\boldsymbol{W}$ is Gaussian such that $\boldsymbol{W} \sim \mathcal{N}(0, \Sigma_W(\theta))$.*

Let us first opt for the exact known form of $\log p(Y, \boldsymbol{W}; \theta)$ but assume that the approximation (3.19) is used in lieu of the analytically intractable $p(\boldsymbol{W}|Y; \theta^{(i)})$.

**Proposition 3.3.4.** *Let the model be subject to Assumption 3.3.3 and assume that Laplace's approximation (3.19) of the posterior $p(\boldsymbol{W}|Y;\theta)$ is used. Then, the intermediate quantity of the EM algorithm, $Q(\theta,\theta^{(i)})$, up to some $\theta$-independent terms is given by*

$$\begin{aligned}\widehat{Q}_1(\theta,\theta^{(i)}) :=& -\frac{1}{2}\log\det(\Sigma_Y(\theta)) - \frac{1}{2}Y^\top\Sigma_Y^{-1}(\theta)Y\\ &-\frac{1}{2}\log\det(\Sigma_W(\theta)) - \frac{1}{2}\mathbb{E}[\mu^\top(\boldsymbol{W};\theta)\Sigma_Y^{-1}(\theta)\mu(\boldsymbol{W};\theta)|Y;\theta^{(i)}]\\ &+Y^\top\Sigma_Y^{-1}(\theta)\mathbb{E}[\mu(\boldsymbol{W};\theta)|Y;\theta^{(i)}] - \frac{1}{2}\mathbb{E}\left[\boldsymbol{W}^\top\Sigma_W^{-1}(\theta)\boldsymbol{W}|Y;\theta^{(i)}\right].\end{aligned} \tag{3.24}$$

*Proof.* By linearity of integrals, (3.8) can be written as a sum of two integrals,

$$\int \log p(Y|W;\theta)\tilde{p}(W|Y;\theta^{(i)})\,\mathrm{d}W + \int \log p(W;\theta)\tilde{p}(W|Y;\theta^{(i)})\,\mathrm{d}W \tag{3.25}$$

in which the posterior of $\boldsymbol{W}$ is approximated by (3.19). According to Assumption 3.3.3, the first factors of the integrands are

$$\begin{aligned}\log p(Y|W;\theta) =& -\frac{d_yN}{2}\log(2\pi) - \frac{1}{2}\log\det(\Sigma_Y(\theta))\\ &-\frac{1}{2}\left(Y^\top\Sigma_Y^{-1}(\theta)Y - 2Y^\top\Sigma_Y^{-1}(\theta)\mu(W;\theta) + \mu^\top(W;\theta)\Sigma_Y^{-1}(\theta)\mu(W;\theta)\right)\end{aligned} \tag{3.26}$$

and

$$\log p(W;\theta) = -\frac{d_wN}{2}\log(2\pi) - \frac{1}{2}\log\det(\Sigma_W(\theta)) - \frac{1}{2}W^\top\Sigma_W^{-1}(\theta)W. \tag{3.27}$$

Ignoring all $\theta$-independent terms in (3.26) and (3.27) and using linearity of integrals once more, we find that (3.25) is

$$\begin{aligned}&\frac{1}{2}\log\det(\Sigma_Y(\theta)) - \frac{1}{2}Y^\top\Sigma_Y^{-1}(\theta)Y - \frac{1}{2}\mathbb{E}[\mu^\top(\boldsymbol{W};\theta)\Sigma_Y^{-1}(\theta)\mu(\boldsymbol{W};\theta)|Y;\theta^{(i)}]\\ &-\frac{1}{2}\log\det(\Sigma_W(\theta)) + Y^\top\Sigma_Y^{-1}(\theta)\mathbb{E}[\mu(\boldsymbol{W};\theta)|Y;\theta^{(i)}] - \frac{1}{2}\mathbb{E}\left[\boldsymbol{W}^\top\Sigma_W^{-1}(\theta)\boldsymbol{W}|Y;\theta^{(i)}\right]\end{aligned}$$

in which all the expectations are with respect to the Gaussian approximation (3.19) with $\theta = \theta^{(i)}$. ■

The expectations in (3.24) are easy to evaluate only in few cases; for example, when the outputs are independent and the entries $[\mu(W;\theta)]_i$ of the mean vector of $\boldsymbol{Y}|W$ are polynomials in $\boldsymbol{w}_t$, see Example 3.3.3. Unfortunately, for the general model (2.10), the first two expectations are again analytically intractable and further approximations of these conditional moments are required (using Monte Carlo simulations for example). The last expectation can be evaluated by observing that

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{W}^\top\Sigma_W^{-1}(\theta)\boldsymbol{W}|Y;\theta^{(i)}\right] &= \mathbb{E}\left[\mathbf{tr}\left[\Sigma_W^{-1}(\theta)\boldsymbol{W}\boldsymbol{W}^\top\right]|Y;\theta^{(i)}\right]\\ &= \mathbf{tr}\left[\mathbb{E}\left[\Sigma_W^{-1}(\theta)\boldsymbol{W}\boldsymbol{W}^\top|Y;\theta^{(i)}\right]\right]\\ &= \mathbf{tr}\left[\Sigma_W^{-1}(\theta)(\Sigma(\theta^{(i)}) + \widehat{W}(\theta^{(i)})\widehat{W}^\top(\theta^{(i)}))\right]\end{aligned} \tag{3.28}$$

where $\Sigma(\theta^{(i)}) := \Sigma(\widehat{W}(\theta^{(i)}), \theta^{(i)})$, and $\mathbf{tr}$ is used to denote the trace operator. In the first equality, we used the fact that the trace of a product of two or more matrices is invariant under cyclic permutations, and in the second equality we used the fact that the trace is a linear operator, and therefore it commutes with the expectation operator. For a proof of these properties, we refer the reader to any book on matrix analysis, see for example [62].

An alternative idea is to use approximations for both $\log p(Y, \boldsymbol{W}; \theta)$ and $p(\boldsymbol{W}|Y; \theta^{(i)})$ in (3.8).

**Proposition 3.3.5.** *Assume that $p(Y, \boldsymbol{W}; \theta)$ is approximated by Laplace's approximation (3.23). Then, the intermediate quantity of the EM algorithm, $Q(\theta, \theta^{(i)})$, up to some $\theta$-independent terms is given by*

$$\begin{aligned}\widehat{Q}_2(\theta, \theta^{(i)}) := {} & \log \tilde{p}(Y, \widehat{W}(\theta); \theta) \\ & - \frac{1}{2}\widehat{W}^\top(\theta)\Sigma^{-1}(\theta)\left(\widehat{W}(\theta) - 2\widehat{W}(\theta^{(i)})\right) \\ & - \frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}(\theta)(\Sigma(\theta^{(i)}) + \widehat{W}(\theta^{(i)})\widehat{W}^\top(\theta^{(i)}))\right].\end{aligned} \tag{3.29}$$

*in which $\Sigma(\theta) = \Sigma(\widehat{W}(\theta); \theta)$.*

*Proof.* Observe that according to Laplace's approximation (3.23) it holds that

$$\log \tilde{p}(Y, \boldsymbol{W}; \theta) = \log \tilde{p}(Y; \theta) + \log \tilde{p}(\boldsymbol{W}|Y; \theta). \tag{3.30}$$

where the first term on the right hand side is independent of $\boldsymbol{W}$. Therefore, by linearity of integrals, (3.8) becomes

$$\log \tilde{p}(Y; \theta) + \int \log \tilde{p}(W|Y; \theta)\tilde{p}(W|Y; \theta^{(i)})\, \mathrm{d}W. \tag{3.31}$$

in which the integral in the second term evaluates to

$$\begin{aligned}& -\frac{d_W N}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma(\theta) - \frac{1}{2}\widehat{W}^\top(\theta)\Sigma^{-1}(\theta^{(i)})\left(\widehat{W}(\theta) - 2\widehat{W}(\theta^{(i)})\right) \\ & - \frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}(\theta)(\Sigma(\theta^{(i)}) + \widehat{W}(\theta^{(i)})\widehat{W}^\top(\theta^{(i)}))\right].\end{aligned}$$

Adding this value to the expression of $\log \tilde{p}(Y; \theta)$ in (3.22) gives (3.31). ■

Notice that defining

$$Q(\theta, \theta^{(i)}) = \widehat{Q}_2(\theta, \theta^{(i)}) = \log \tilde{p}(Y; \theta) + \int \log \tilde{p}(W|Y; \theta)\tilde{p}(W|Y; \theta^{(i)})\, \mathrm{d}W,$$

as suggested in Proposition 3.3.5, corresponds exactly to (3.7) with $p$ replaced by $\tilde{p}$. This means that using $\widehat{Q}_2(\theta, \theta^{(i)})$ in an EM algorithm will generate a sequence $\{\theta^{(i)}\}$ such that $\tilde{p}(Y; \theta^{(i+1)}) \geq \tilde{p}(Y; \theta^{(i)})$ and convergence to a stationary point of the approximate likelihood $\tilde{p}(Y;\theta)$ is guaranteed (see Theorem 3.2.1 and the associated discussion). The E-step and the M-step of such an algorithm are defined by maximization problems over $W$ and $\theta$ respectively and, in general, will not admit closed-form solutions. This means that direct maximization of the approximate log-likelihood $\log \tilde{p}(Y;\theta)$ over $\theta$ using a gradient-based optimization algorithm might be computationally cheaper. On the other hand, as observed in [106, 107, 125], the log-likelihood function can exhibit an erratic behavior with many local maxima creating a difficulty for gradient-based algorithms, and the EM algorithm seems to be able to avoid these local maxima. One of the given explanations is that the intermediate quantity is well-behaved in comparison to the erratic log-likelihood function; however, a thorough understanding of this behavior is still missing and there exist no guarantees that the EM is better than any other local optimization algorithm in avoiding local solutions.

When the approximate likelihood function $\tilde{p}(Y;\theta)$ is straightforward to maximize over $\Theta$, an EM algorithm based on $\widehat{Q}_2(\theta,\theta^{(i)})$ does not seem to have advantage over direct optimization. To clarify this point, we apply the EM algorithm, in the suggested form considering a trivial (static) model.

**Example 3.3.1.** Consider the static model

$$\boldsymbol{y} = \theta u + \boldsymbol{w} + \boldsymbol{v}$$

in which $\theta = 0.7$, $\boldsymbol{w}$ and $\boldsymbol{v}$ are standard Gaussian random variables, and $u$ is a known real number. Assume that we observed only a single sample of $\boldsymbol{y}$ and our goal is to estimate $\theta$ using the MLE. It is trivial to see that

$$p(y;\theta) \propto \exp\left(-\frac{1}{4}(y-\theta u)^2\right)$$

and therefore the ML estimate is $\hat{\theta} = y/u$.

Now observe that, since $\boldsymbol{y}$ and $\boldsymbol{w}$ are jointly Gaussian random variables, the posterior

$$p(w|y;\theta) \propto \exp\left(-(w-(-\frac{1}{2}(y-\theta u))^2\right)$$

and

$$Q(\theta,\theta^{(i)}) = \log p(y;\theta) + \int \log p(w|y;\theta)p(w|y;\theta^{(i)})\,\mathrm{d}w$$

The integral on the right hand side is easy to evaluate and we find that maximizing $Q(\theta,\theta^{(i)})$ over $\theta$ is the same as evaluating

$$\theta^{(i+1)} = \arg\min_{\theta}\ \frac{1}{2}(y-\theta u)^2 + \frac{1}{4}(y-\theta u)(y-\theta^{(i)}u). \tag{3.32}$$

Therefore, the EM iterations based on $Q$ as written in (3.31) reads

$$\theta^{(i+1)} = \frac{y}{u} - \frac{1}{4}\frac{(y-\theta^{(i)}u)}{u}.$$

This is a fixed-point iteration with a limit $\hat{\theta} = y/u$. The point is that, for such a simple model, there is no advantage of iterating over (3.32) instead of directly minimizing the first term in the same equation.

Before describing possible approximations of gradient based algorithms, we summarize the EM algorithm based on $\widehat{Q}_1(\theta, \theta^{(i)})$ in Algorithm 2. The algorithm is first applied to the case of a linear Gaussian model with latent process (as in (2.25) for example), and then we perform a simulation study on a simple nonlinear model.

**Algorithm 2:** An Expectation-Maximization (EM) algorithm based on Laplace's approximation of the posterior.

```
input  : An initial guess $\theta^{(0)}$, the data $(Y, U)$, and a stopping criterion
output : An estimate $\hat{\theta}$
Initialize $\theta^{(i)}$ and set index $i \leftarrow 0$
while not converged do
    Compute $\widehat{W}(\theta^{(i)})$ and $\Sigma_W(\widehat{W}(\theta^{(i)}), \theta^{(i)})$ by solving $\max_W p(Y, W; \theta^{(i)})$.
    Compute $\theta^{(i+1)} \in \arg\max_{\theta \in \Theta} \widehat{Q}_1(\theta, \theta^{(i)})$ (see (3.24))
    $i \leftarrow i + 1$
end
Set $\hat{\theta} = \theta^{(i)}$
```

**Example 3.3.2** (EM based on Laplace's approximation for linear Gaussian models)**.** Consider a linear Gaussian model described by

$$\boldsymbol{Y} = F(\theta)\boldsymbol{W} + \boldsymbol{V} \tag{3.33}$$

such that

$$\boldsymbol{W} \sim p(\boldsymbol{W}; \theta) = \mathcal{N}(0, \Sigma_W(\theta)), \quad \text{and} \quad \boldsymbol{V} \sim p(\boldsymbol{V}; \theta) = \mathcal{N}(0, \Sigma_V(\theta)).$$

For clarity of exposition, we assume that there is no input ($U = 0$). The general case with either stochastic (observed) or deterministic input $U \neq 0$ is not different if we assume that $U$ is generated independently of both $\boldsymbol{W}$ and $\boldsymbol{V}$. The only changes will appear in the mean of the distributions.

Using standard results for multivariate Gaussian random variables (see Appendix C), it is easy to see that $\boldsymbol{Y}$ and $\boldsymbol{W}$ are jointly Gaussian such that

$$\begin{aligned} \boldsymbol{Y} &\sim p(\boldsymbol{Y}; \theta) = \mathcal{N}(0, \Sigma_Y(\theta)), \\ \boldsymbol{W}|Y &\sim p(\boldsymbol{W}|Y; \theta) = \mathcal{N}(\widehat{W}_{\text{MAP}}, \Sigma_{W|Y}(\theta)), \end{aligned}$$

in which

$$\begin{aligned} \Sigma_Y(\theta) &= F(\theta)\Sigma_W(\theta)F^\top(\theta) + \Sigma_V(\theta), \\ \widehat{W}_{\text{MAP}} &= \Sigma_W(\theta)F^\top(\theta)\Sigma_Y^{-1}(\theta)Y, \text{ and} \\ \Sigma_{W|Y}(\theta) &= \Sigma_W(\theta) - \Sigma_W(\theta)F^\top(\theta)\Sigma_Y^{-1}(\theta)F(\theta)\Sigma_W(\theta). \end{aligned} \tag{3.34}$$

Laplace's approximation method requires the computation of (3.17). For the given model in (3.33), it is easy to see that (note that $\log p(Y,W;\theta)$ has the same stationary points as $p(Y,W;\theta)$)

$$\begin{aligned}\log p(Y,W;\theta) &\propto -\frac{1}{2}(Y-F(\theta)W)^{\top}\Sigma_V^{-1}(\theta)(Y-F(\theta)W)-\frac{1}{2}W^{\top}\Sigma_W^{-1}(\theta)W\\ &\propto Y^{\top}\Sigma_V^{-1}(\theta)F(\theta)W-\frac{1}{2}W^{\top}\left(\Sigma_W^{-1}(\theta)-F^{\top}(\theta)\Sigma_V^{-1}(\theta)F(\theta)\right)W\end{aligned}$$

which is quadratic in $W$ and therefore has a unique (global) maximizer. To find it, we compute

$$\nabla_W \log p(Y,W;\theta)\propto Y^{\top}\Sigma_V^{-1}(\theta)F(\theta)-W^{\top}\left(\Sigma_W^{-1}(\theta)-F^{\top}(\theta)\Sigma_V^{-1}(\theta)F(\theta)\right)^{-1},$$

equate it to zero and solve for $W$; we get that

$$\widehat{W}_{\mathrm{MAP}}(\theta)=\left(\Sigma_W^{-1}(\theta)-F^{\top}(\theta)\Sigma_V^{-1}(\theta)F(\theta)\right)^{-1}F^{\top}(\theta)\Sigma_V^{-1}(\theta)Y. \tag{3.35}$$

Using the Woodbury formula (see [62, Section 0.7.4]) and a couple of properties of matrix inverses, it can be shown that the right-hand side of (3.35) and that of the second row of (3.34) are identical.

To find the covariance matrix of Laplace's approximation, we evaluate the Hessian matrix of $\log p(Y,W;\theta)$ and observe that

$$\nabla_W^2 \log p(Y,W;\theta)\propto -\left(\Sigma_W^{-1}(\theta)-F^{\top}(\theta)\Sigma_V^{-1}(\theta)F(\theta)\right) \tag{3.36}$$

which is independent of $W$. A direct application of the Woodbury formula shows that the negative inverse of (3.36) coincides with the posterior covariance in (3.34).

Therefore, Laplace's approximation is in fact exact for linear Gaussian models and $p(\boldsymbol{W}|Y;\theta)=\tilde{p}(\boldsymbol{W}|Y;\theta)$. This also means that Laplace's approximation of the normalizing constant (3.21) coincides with the true likelihood function; $p(Y;\theta)=\tilde{p}(Y;\theta)$. Consequently, Algorithm 1 and Algorithm 2 are equivalent for linear Gaussian models.

Furthermore, the expectations in (3.24) are available in closed-form:

$$\begin{aligned}\mathbb{E}_W[F(\theta)W|Y;\theta^{(i)}]&=F(\theta)\widehat{W}(\theta^{(i)}),\\ \mathbb{E}_W[W^{\top}F^{\top}(\theta)\Sigma_Y^{-1}(\theta)F(\theta)W|Y;\theta^{(i)}]&=\\ \mathbf{tr}\left[\Sigma_Y^{-1}(\theta)F(\theta^{(i)})\left(\Sigma_W(\theta^{(i)})+\widehat{W}(\theta^{(i)})\widehat{W}^{\top}(\theta^{(i)})\right)F^{\top}(\theta^{(i)})\right]&.\end{aligned}$$

In the next example, Algorithm 2 is tested on a nonlinear model with bi-modal posterior using numerical simulations.

**Example 3.3.3** (Stochastic Wiener model with quadratic nonlinearity)**.** Assume that the data is generated according to the model

$$\begin{aligned} \boldsymbol{x}_t &= \frac{q^{-1}}{1-\theta q^{-1}} u_t + \boldsymbol{w}_t \\ \boldsymbol{y}_t &= \boldsymbol{x}_t^2 + \boldsymbol{v}_t, \quad t = 1, \ldots, N, \end{aligned} \tag{3.37}$$

in which all the signals are scalar (first order SISO model), $\theta = 0.7$, $\boldsymbol{w}_t \sim \mathcal{N}(0, \lambda_w)$, and $\boldsymbol{v}_t \sim \mathcal{N}(0,1)$. The input $u_t$ is a known realization of $\boldsymbol{u}_t \sim \mathcal{N}(0,1)$. Observe that the posterior of $\boldsymbol{W}$ for this model can be bi-modal due to the quadratic nonlinearity in the output equation. We assume that $\boldsymbol{w}$ and $\boldsymbol{v}$ are independent and mutually independent; thus, the outputs are independent over time and it is possible to approximate the likelihood function arbitrarily well using numerical integration methods (see Section 3.4). Such an approximation can be used within a quasi-Newton algorithm to compute the true MLE to an arbitrary accuracy. In this example, Gauss-Hermite quadrature (see [37, Section 5.3.4]) is used to compare the true MLE to the estimate of Algorithm 2. Due to the independence of the outputs, all the expectations defining the intermediate quantity in (3.24) can be computed in closed form (as functions of the mean and the variance of (3.19)).

The results of the two algorithms starting from the initial value 0.4 are shown in Figure 3.3 for $N = 100$ and the following values for $\lambda_w : 0.1, 0.3,$ and 0.6 respectively from left to right.

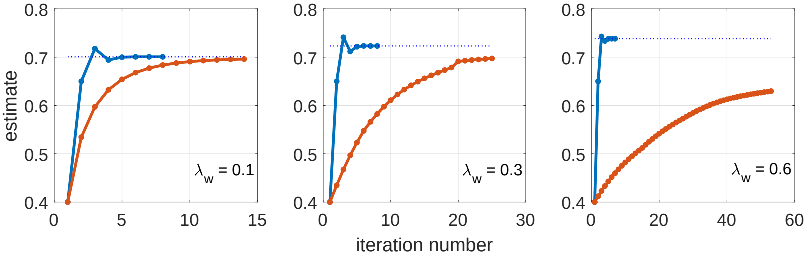


**Figure 3.3:** The quasi-Newton iterations solving the true ML problem are given by the blue curves, while the iterations of Algorithm 2 are shown in red. The three cases from left to right are with $\lambda_w = 0.1, 0.3,$ and 0.6 respectively.

As can be seen from this simulation result, the approximation (for this example) is acceptable for small variances $\lambda_w$, however when $\lambda_w$ is increased the approximation is not reliable anymore. We also observe that it takes more iterations for the algorithm to converge. The stopping criterion for this example was $|\theta^{(i)} - \theta^{(i-1)}| < 10^{-3}$.

These simulation results do not improve if the model is such that the posterior is uni-modal. Assume for example that the measurement model is given by

$$\boldsymbol{y}_t = \boldsymbol{x}_t^3 + \boldsymbol{v}_t \tag{3.38}$$

instead of the quadratic model in (3.37). In this case, the posterior of $\boldsymbol{W}$ is expected to have one dominant mode. However, if we repeat the same simulation experiment as above, we get the results shown in Figure 3.4. The approximations are actually worse and the algorithm takes longer to converge.

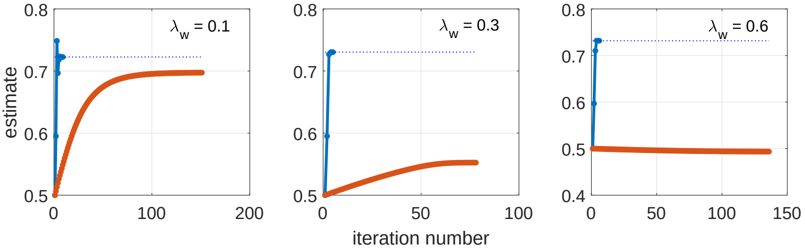


**Figure 3.4:** The results for the same cases as in Figure 3.3 however for a cubic measurement model (3.38)

Even though the example is simple enough to allow for an accurate approximation of the MLE by using numerical integration, Algorithm 2 does not seem very promising and it is difficult to assess its convergence.

An alternative idea is to use Laplace's approximation method to directly approximate the intermediate quantity, instead of first approximating the posterior of $\boldsymbol{W}$ and then evaluating the involved expectations. For example, an approximation can be obtained by assuming that only one point has the major contribution to the value of the integral. We may define the approximation

$$Q(\theta, \theta^{(i)}) \approx \widehat{Q}_3(\theta, \theta^{(i)}) := \log p(Y, \widehat{W}(\theta^{(i)}); \theta)$$

in which $\widehat{W}(\theta^{(i)})$ is used. Essentially, this means that we are assuming that the posterior of $\boldsymbol{W}$ (or the approximation (3.19)) is concentrated around its mean value (i.e., the covariance is very close to zero). We summarize this idea in Algorithm 3 in which the E-step is replaced by a maximization problem over $W$. Hence, the algorithm iterates between two maximization problems and it is easy to see that it is solving the following "joint" estimation problem

$$(\widehat{W}, \hat{\theta}) := \arg\max_{W,\theta} \; \log p(Y, W; \theta) \tag{3.39}$$

by maximizing over one variable at a time (i.e., coordinate ascent). This means that the iterations of an EM algorithm based on the intermediate value $\widehat{Q}_3(\theta, \theta^{(i)})$ will converge to some value.

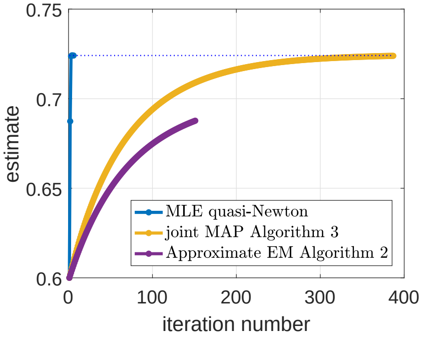


**Figure 3.5:** Comparison of the MLE, joint MAP estimate as given by Algorithm 3 , and the approximate EM algorithm 2

Using an EM algorithm to solve (3.39) would be preferred if maximizing over $W$ or $\theta$ individually is a tractable problem with a closed-form solution. However, in general, using a quasi-Newton algorithm to directly solve (3.39) is expected to be faster than Algorithm 3 due to the intractability of the optimization problems over $\theta$ and $W$.

The idea of joint estimation is quite old (see [25] for example) and it is known that this estimator (as shown for the linear case in [63]) is not consistent and the resulting estimator of $\theta$ does not coincide with the MLE. On the other hand, it has been noticed (see [150]) that for short data sizes the estimate of $\theta$ given in (3.39) could be preferable to the MLE in terms of both the bias and the covariance.

**Example 3.3.4** (Joint MAP estimation)**.** In this example, we compare Algorithm 2 to Algorithm 3 using the model (3.37) introduced in Example 3.3.3. For the current example, we have

$$\lambda_w = \lambda_v = 1, \ N = 100, \text{ and } \ \boldsymbol{u}_t \sim \mathcal{N}(1,1).$$

The algorithms are terminated once $\|\theta^{(i)} - \theta^{(i-1)}\| < 10^{-5}$. The simulation results are shown in Figure 3.5. It shows that the joint MAP estimate comes very close to the MLE (very slowly with many iterations), while Algorithm 2 still shows the same speculative behavior observed in Example 3.3.3.

Before moving to the next part, we have the following lemma regarding the intermediate quantity. It is interesting to observe that in order to approximate the iterations of the EM algorithm (see (3.10)), we do not actually need to evaluate the value of the $Q$-function itself; it is sufficient to evaluate $Q$ up to a $\theta$-independent factor. As a result, the quantity to be maximized in the M-step can be written as an expectation with respect to the known prior PDF $p(\boldsymbol{W};\theta)$, and therefore (subject to Assumption 3.1) the integrand is known. This fact will be of interest when discussing possible Monte Carlo approximations in the next sections.

**Algorithm 3:** An Expectation-Maximization (EM) algorithm based on a degenerate approximation of the posterior (joint MAP estimation)

```
input  : An initial guess $\theta^{(0)}$, the data $(Y, U)$, and a stopping criterion
output : An estimate $\hat{\theta}$
Initialize $\theta^{(i)}$ and set index $i \leftarrow 0$
while not converged do
    Compute $\widehat{W}(\theta^{(i)})$ by solving $\max_W p(Y, W; \theta^{(i)})$.
    Compute $\theta^{(i+1)} \in \arg\max_{\theta \in \Theta} \widehat{Q}_3(\theta, \theta^{(i)})$, where
            $\widehat{Q}_3(\theta, \theta^{(i)}) := \log p(Y, \widehat{W}(\theta^{(i)}); \theta)$
    $i \leftarrow i + 1$
end
Set $\hat{\theta} = \theta^{(i)}$
```

**Lemma 3.3.6.** *The intermediate quantity of the EM algorithm satisfies*

$$Q(\theta, \theta^{(i)}) \propto \mathbb{E}[p(Y|\boldsymbol{W}; \theta^{(i)}) \log p(Y, \boldsymbol{W}; \theta); \theta^{(i)}], \tag{3.40}$$

*where the expectation is with respect to the known prior PDF* $p(\boldsymbol{W}; \theta^{(i)})$*.*

*Proof.* Note that according to the expression in (3.8), it holds in general that

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= \int \log p(Y, W; \theta) \frac{p(Y, W; \theta^{(i)})}{p(Y; \theta^{(i)})} \, \mathrm{d}W \\ \{p(Y; \theta^{(i)}) \text{ independent of } W\} &= \frac{1}{p(Y; \theta^{(i)})} \int \log p(Y, W; \theta) p(Y, W; \theta^{(i)}) \, \mathrm{d}W \\ \{\text{as a function of } \theta\} &\propto \int \log p(Y, W; \theta) p(Y, W; \theta^{(i)}) \, \mathrm{d}W \\ &= \int \log p(Y, W; \theta) \; p(Y|W; \theta^{(i)}) p(W; \theta^{(i)}) \, \mathrm{d}W \\ &= \mathbb{E}\left[\log p(Y, \boldsymbol{W}; \theta) p(Y|\boldsymbol{W}; \theta^{(i)}) \; ; \theta^{(i)}\right]. \end{aligned}$$

■

If the expectation in (3.40) can be evaluated analytically, the E-step is tractable. Otherwise, approximations are necessary. It should be noted that, even though the integrand in (3.40) is known exactly, evaluating the expectation is not an easy task. For example, let the model be subject to Assumption 3.3.3 and assume that the model parameterization is such that the prior $p(\boldsymbol{W})$ is $\theta$-independent. In this case, (3.40) becomes

$$\begin{aligned} &\mathbb{E}[p(Y|\boldsymbol{W}; \theta^{(i)}) \log p(Y, \boldsymbol{W}; \theta); \theta^{(i)}] \\ &\qquad = \mathbb{E}[p(Y|\boldsymbol{W}; \theta^{(i)}) \log p(Y|\boldsymbol{W}; \theta)] + \mathbb{E}[p(Y|\boldsymbol{W}; \theta^{(i)}) \log p(\boldsymbol{W})] \end{aligned}$$

and since the second term of the right hand side is $\theta$-independent,

$$Q(\theta, \theta^{(i)}) \propto \mathbb{E}[p(Y|\boldsymbol{W}; \theta^{(i)}) \log p(Y|\boldsymbol{W}; \theta)].$$

Using the expression of the conditional likelihood in (3.26) and ignoring all terms that are $\theta$-independent, we get that

$$\begin{aligned} Q(\theta, \theta^{(i)}) \propto & -\frac{1}{2} \log \det(\Sigma_Y(\theta)) - \frac{1}{2} Y^\intercal \Sigma_Y^{-1}(\theta) Y \\ & + Y^\intercal \Sigma_Y^{-1}(\theta) \mathbb{E}[\mu(\boldsymbol{W}; \theta) \varphi(Y, \boldsymbol{W}; \theta^{(i)})] \\ & - \frac{1}{2} \mathbb{E}[\mu^\intercal(\boldsymbol{W}; \theta) \Sigma_Y^{-1}(\theta) \mu(\boldsymbol{W}; \theta) \varphi(Y, \boldsymbol{W}; \theta^{(i)})] \end{aligned} \tag{3.41}$$

in which

$$\varphi(Y, \boldsymbol{W}; \theta^{(i)}) := \exp\left(\left(Y^\intercal - \frac{1}{2} \mu^\intercal(\boldsymbol{W}; \theta^{(i)})\right) \Sigma_Y^{-1}(\theta^{(i)}) \mu(\boldsymbol{W}; \theta^{(i)})\right).$$

Thus, the two expectations in (3.41) are, in general, analytically intractable. Furthermore, it will turn out that, for any reasonable value of $N$, naïve Monte Carlo approximations based on direct samplings are extremely inefficient and come with very large variance (see Example 3.4.2).

**Remark 3.3.7** (Analytic approximations of the EM algorithm)**.**

- *Laplace's approximation is one way of obtaining a Gaussian approximation of the posterior PDF $p(\boldsymbol{W}|Y; \theta)$ and an approximate value of the likelihood function. As shown above, with such approximations, the E-step of the EM algorithm simplifies and it is possible to obtain closed-form expressions for the intermediate quantity if the expectations are tractable. The idea of using Gaussian approximations is not specific to Laplace's method or the EM algorithm. It is possible to use alternative techniques to obtain Gaussian approximations, and we shall see this later in the context of PEMs in the following chapter.*

- *An alternative analytical approximation for multi-modal posteriors may be obtained using a Gaussian-Mixture PDF instead of a single Gaussian. This can be seen as an extension of the EM algorithm based on Laplace's approximation to an EM algorithm based on Gaussian-Mixture approximations. Gaussian-Mixture PDFs can be constructed, for example, based on Laplace's approximations at different modes; however, we will not pursue this idea further.*

In the coming section, we discuss a possible use of Laplace's approximation within a quasi-Newton algorithm. Given the data and a candidate $\theta^{(i)}$, the idea is to use Laplace's approximation method to compute an approximate value for the likelihood function (and possibly its gradient). This idea seems better when compared to Algorithm 2. To make an argument for this, observe that in Algorithm

2, only the posterior was approximated and the true (assumed known) joint PDF $p(\boldsymbol{Y}, \boldsymbol{W}; \theta)$ was used. The consequence is that all the identities of the EM algorithm, as shown in Section 3.2.1, are not valid anymore and it is not obvious how to establish the convergence of the algorithm (to any point). On the other hand, given a well defined function that can be evaluated at every candidate point, and under some smoothness conditions, the convergence of gradient-based algorithms to a limit point can be guaranteed. It is a different issue then whether the resulting approximation is consistent or not.

### 3.3.2 Approximate Gradient-based Methods

Laplace's approximation method suggests the value in (3.21) as an approximation of the likelihood function. It can be used to approximate the MLE by computing

$$\arg\max_{\theta} \; \tilde{p}(Y;\theta) = \arg\max_{\theta} \; \log \tilde{p}(Y;\theta).$$

Ignoring all $\theta$-independent terms, this is equal to

$$\arg\max_{\theta} \; \log p(Y, \widehat{W}(\theta); \theta) + \frac{1}{2} \log\det(\Sigma_W(\widehat{W}(\theta), \theta)), \tag{3.42}$$

and it is clear that the approximation is based on joint maximizations over $W$ and $\theta$ which is different from (3.39). One can think of (3.42) as a regularized joint MAP estimate. Observe that this estimator is consistent in the linear case under mild conditions (it coincides with the MLE), unlike (3.39). To see this, we borrow an argument used in [63, Appendix A]: let us write

$$(\widehat{W}, \hat{\theta}) = \arg\max_{W,\theta} \; p(W, \theta|Y) = \arg\max_{W,\theta} \; p(Y, W|\theta)p(\theta)$$

in which $p(\theta)$ is a correction (regularization) factor which is independent of $W$. For a fixed value $\theta$, the maximizer $\widehat{W}(\theta)$ is the MAP estimate of $\boldsymbol{W}$ because maximizing $p(Y, W|\theta)$ over $W$ is the same as maximizing $p(W|Y, \theta)$. On the other hand, the estimator of the parameters which is defined by

$$\hat{\theta} = \arg\max_{\theta} \; p(Y, \widehat{W}(\theta)|\theta)p(\theta)$$

is not consistent in general[5] and does not coincide with the MLE. Observe that we can rewrite the estimator in the form

$$\hat{\theta} = \arg\max_{\theta} \; \underbrace{p(\widehat{W}(\theta)|Y, \theta)p(Y|\theta)}_{=p(Y, \widehat{W}(\theta)|\theta)} p(\theta) \tag{3.43}$$

which indicates that the maximizer will coincide with the MLE if and only if $p(\widehat{W}(\theta)|Y, \theta)p(\theta)$ is not a function of $\theta$ (see [63]).

[5] For consistency, $p(\theta)$ has to be such that, in the limit as $N \to \infty$ and under some regularity conditions, $\theta^\circ$ is the unique minimizer of the averaged objective function.

This can be seen as another motivation for the cost function in (3.42). If the posterior is approximated by the multivariate Gaussian PDF in (3.19), we can find an approximation of $p(\theta)$ if we assume that $p(\widehat{W}(\theta)|Y,\theta)$ is approximated by the normalization factor in (3.19) and define $p(\theta)$ as its inverse. We then use this approximate correction term in (3.43) together with the true joint model $p(Y,\widehat{W}(\theta)|\theta)$.

This point of view shows that, regardless of the posterior, we can always write the maximum likelihood estimate as

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \ \log p(Y,\widehat{W}(\theta);\theta) - \log p(\widehat{W}(\theta)|Y;\theta)\\ &= \arg\max_{\theta} \ p(Y;\theta)\end{aligned}$$

which comes in agreement with the comments after (3.4). Consequently, even in cases with non-unique MAP estimates $\widehat{W}_{MAP}(\theta)$, any of them can be used assuming that we are able to compute $p(\widehat{W}(\theta)|Y;\theta)$ (which depends on the unknown likelihood). Assuming a case where we are able to reasonably approximate the posterior around one mode by a multivariate Gaussian, an approximation of $\log p(\widehat{W}(\theta)|Y;\theta)$ can be obtained. The quality of the estimates will however depend on the shape of the true posterior and will depend on the used input signal. To clarify this point, we consider the following example.

**Example 3.3.5** (Bi-modal posteriors)**.** Consider the following static nonlinear model

$$\boldsymbol{y} = (u+\boldsymbol{w})^r + \boldsymbol{v} \tag{3.44}$$

in which $u$ is a given real number, $\boldsymbol{w}$ and $\boldsymbol{v}$ are standard Gaussian (scalar) random variables. Assume that we observed a single sample of $\boldsymbol{y}$. Then it is possible to compute the true posterior of $\boldsymbol{w}$, which is given by

$$p(\boldsymbol{w}|y,u) = \frac{p(y|u,\boldsymbol{w})p(\boldsymbol{w})}{\int p(y|u,w)p(w)\,\mathrm{d}w},$$

by solving the scalar integral in the denominator (using deterministic numerical integration for example).

Figures 3.6 and 3.7 compare the true posterior of $\boldsymbol{w}$ to the Gaussian approximation obtained by Laplace's approximation method for two cases: when $r = 2$ and when $r = 3$ respectively. Observe that we used different scales for the true and the approximate density plots. In each case, we show the results for the values $u = 0, 1,$ and $5$.

With $r = 2$, it is clear that when $u = 0$ the posterior of $\boldsymbol{w}$ can have two equivalent peaks, as the used realization shows, and the PDF is symmetrical around the vertical axis. Each mode can be seen as a Gaussian PDF by itself and the true posterior looks like a Gaussian mixture with equal mixture weights. In this case, Laplace's approximation is able to capture one of the modes (depending

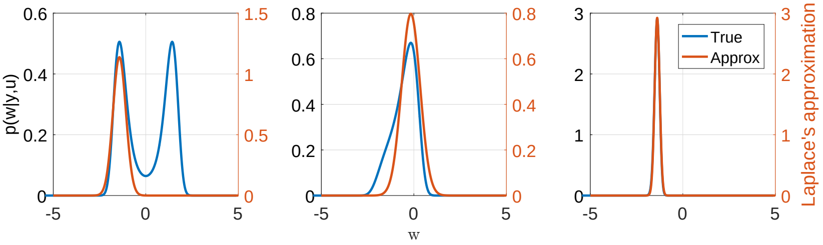


**Figure 3.6:** True and approximate posterior of $\boldsymbol{w}$ for the model in (3.44) when $r = 2$ for different values of $u$. From left to right $u = 0, 1$, and 5.

on the sign of the initialization value of (3.17)). The true value of $p(y)$ for the used realization is 0.1 and Laplace's approximation gives 0.0447 which is about half the original probability. This can be seen from the figure by noticing that $\tilde{p}(\hat{w}|y,u)$ is approximately twice the true value. When $u$ is increased, see the middle and right plots in Figure 3.6, we see that the distribution becomes unimodal and the obtained approximation is quite good. When $u = 5$ the posterior is almost a Gaussian and the approximate value of $p(y)$ is practically the same as the true value. Similar observations are seen when $r = 3$ as shown in Figure 3.7.

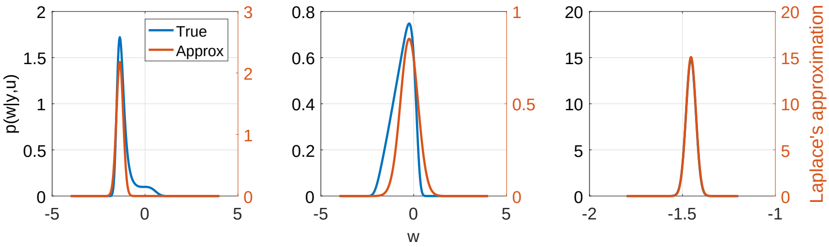


**Figure 3.7:** True and approximate posterior of $\boldsymbol{w}$ for the model in (3.44) when $r = 3$ for different values of $u$. From left to right $u = 0, 1$, and 5.

---

The next example demonstrates the approximation of the likelihood function of a nonlinear model when the posterior is bi-modal. We look at two cases: a case when the model has no inputs, and a second case when there is a small input.

---

**Example 3.3.6** (Likelihood approximation of a bi-modal model)**.** Consider the model

$$\boldsymbol{y}_t = (\theta \boldsymbol{w}_t)^2 + \boldsymbol{v}_t, \quad t = 1, \ldots, N, \tag{3.45}$$

in which $\boldsymbol{w}_t \sim \mathcal{N}(0, 2)$, $\boldsymbol{v}_t \sim \mathcal{N}(0, 0.1)$ are independent over $t$ and mutually independent, and let $\theta = 0.5$. We fixed $N = 100$ and simulated one realization of the data. The true negative log-likelihood and the true posterior of $\boldsymbol{w}_t$ are

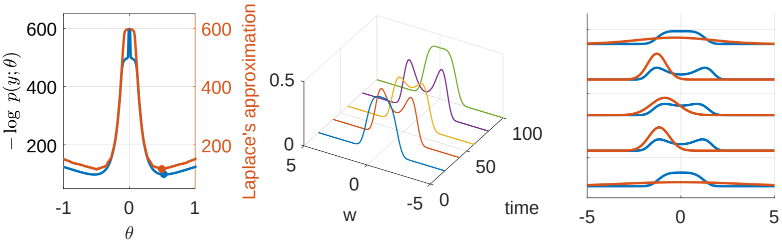


**Figure 3.8:** On the left: The true negative log-likelihood function (blue) and its approximation (red). In the middle: The posterior of $w_t$ at five selected time points. On the right: The approximation of the posterior (red) in comparison to the true one.

then compared to the approximation obtained by Laplace's approximation method. The results are shown in Figure 3.8. It is clear that the posterior can be bi-modal as shown in the middle plot. The right plot shows how Laplace's approximation tries to capture one of the two modes (shown in red). Despite the poor approximation, the shape of the negative log-likelihood can be approximated quite well as shown on the left. Both the approximation and the true negative log-likelihood have very close minima around ±0.5. Observe that the model is not globally identifiable, and the likelihood function has two global maximizers, see Figure 3.9.

In cases where the posterior PDF has different values at each mode, for example when we introduce a small input such that

$$\boldsymbol{y}_t = (u_{t-1} + \theta \boldsymbol{w}_t)^2 + \boldsymbol{v}_t, \tag{3.46}$$

$u_t$ is a known realization of $\boldsymbol{u}_t \sim \mathcal{N}(0.1, 1)$, the peaks of the likelihood function approximation will not be equal. According to which posterior mode (the positive or the negative) is captured by Laplace's approximation, the peak with the corresponding sign will be larger, see Figure 3.10. However, because the model is not identifiable, we cannot hope to recover the sign of $\theta$. The important observation is that the approximation has the same shape as the likelihood function and appears to indicate the location of the minima.

Regrettably, the approximate likelihood function (3.42) is a complicated function of $\theta$ due to the complicated parameterization via $\widehat{W}(\theta)$ and the corresponding covariance matrix $\Sigma_W(\widehat{W}(\theta), \theta)$. If $\widehat{W}(\theta)$ does not assume a closed-form expression, it will not be possible to compute the gradient with respect to $\theta$ analytically and the objective function $\log \tilde{p}(Y;\theta)$ will not be available in closed-form. Nevertheless, for every given candidate $\theta^{(i)} \in \Theta$, it is possible to evaluate the value $\log \tilde{p}(Y;\theta^{(i)})$ by solving the smoothing problem (3.17).

Fortunately, being able to evaluate $\log \tilde{p}(Y;\theta)$ for every given value $Y$ and $\theta \in \Theta$ would be enough for several available numerical optimization algorithms. For example a gradient-free algorithm like the Nelder-Mead algorithm would work. However, there are no guarantees that the Nelder-Mead algorithm converges to

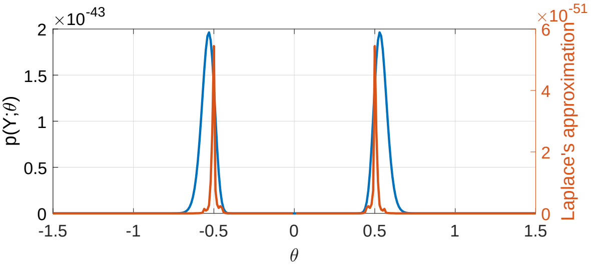


**Figure 3.9:** The true likelihood function (blue) and its approximation (red) for the model in (3.45).

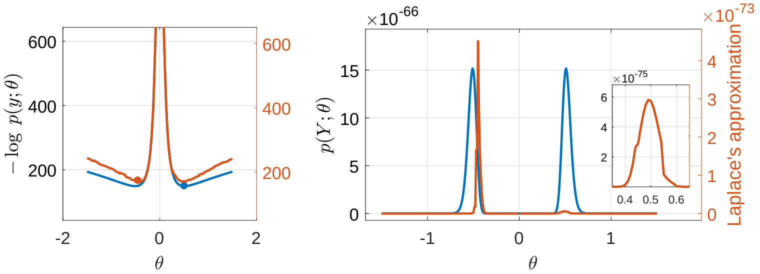


**Figure 3.10:** The negative log-likelihood (left) and the likelihood function (right) for the model in (3.46). The true value is in blue and its approximation is in red. The peak of the approximate likelihood function around 0.5 is rescaled in a subplot.

a local solution (even if the objective function is convex, see for example [75]). Another possibility that could be preferred (due to convergence properties) is the use of the function evaluations to approximate the gradient numerically by using finite differences, see [108, Chapter 8] or [4, Chapter VII], and then use these approximations within a quasi-Newton algorithm or similar gradient-based methods. Most of the optimization toolboxes in modern scientific computing packages like MATLAB (`fminunc`), Mathematica (`FindMinimum`) or R (`optim`) include generic well-implemented functions that apply this strategy.

In the next example, we will use a quasi-Newton algorithm based on approximate gradients to find the ML estimate and its approximation based on (3.42).

**Example 3.3.7** ( Laplace's approximation of the likelihood function)**.** Consider the Wiener model (3.37) introduced in Example 3.3.3. Recall that for this simple model, it is possible to compute the likelihood function by solving a numerical integration problem. In the current example, we would like to compare the approximate likelihood in (3.42) to the true likelihood.

We evaluated the solution to (3.42) under different inputs when $N = 100$, $\boldsymbol{w}_t \sim \mathcal{N}(0,2)$ and $\boldsymbol{v}_t \sim \mathcal{N}(0,1)$. Figures 3.11, 3.12 and 3.13 show the obtained results for three cases when $\boldsymbol{u}_t \sim \mathcal{N}(1,1)$, $\boldsymbol{u}_t \sim \mathcal{N}(0,1)$, and $\boldsymbol{u}_t \sim \mathcal{N}(0,10)$

respectively. Recall that the true value of $\theta$ is 0.7.

For the first case, as shown on the right in Figure 3.11, the shape of the approximate cost function is very close to the true likelihood function, and more interestingly they have very close minima. Running a quasi-Newton algorithm demonstrates the fast convergence of the Laplace's approximate estimate which, as shown on the left in Figure 3.11, is very close to the true ML estimate.

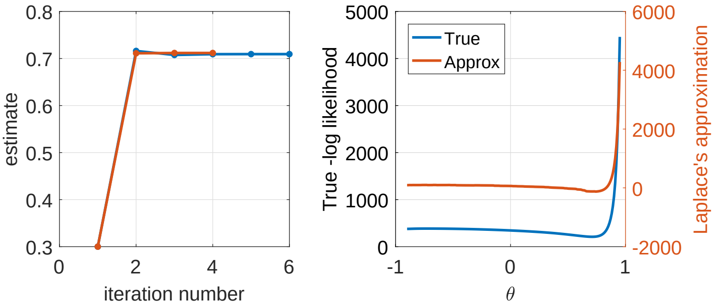


**Figure 3.11:** $\lambda_u = \lambda_v = 1$ and $\lambda_w = 2$, the mean value of $u$ is 1.

However, as shown in Figure 3.12, when we set the mean of the input to zero and keep the same variance, the approximate cost function becomes erratic with several local minima. Yet, the global minimum seem to be close to the true ones.

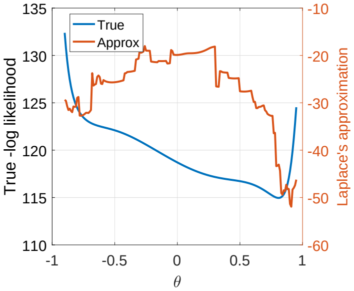


**Figure 3.12:** $\lambda_u = \lambda_v = 1$ and $\lambda_w = 2$, the mean value of $u$ is 0.

Finally, keeping a zero mean value for the input but increasing its variance to 10 seems to improve the situation. As shown in Figure 3.13, the approximate cost function recovers the shape of the true likelihood function, and the estimate is again close to the ML estimate.

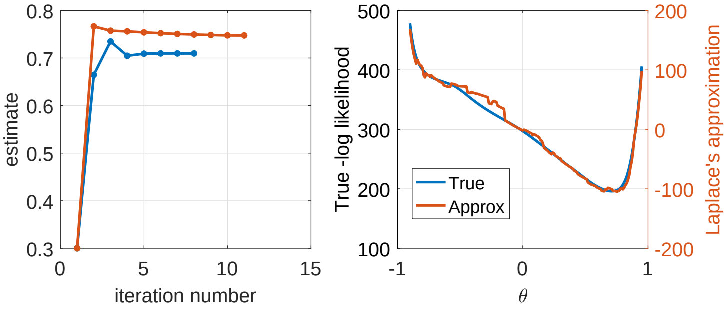


**Figure 3.13:** $\lambda_u = 10, \lambda_v = 1$ and $\lambda_w = 2$, the mean value of $u$ is 0.

### Gradient approximations

A well known tool usually used to evaluate the gradient of the log-likelihood function when the likelihood function is written as a marginalization integral is Fisher's identity. Under some regularity conditions that allow changing the order of differentiation and integration (see [98, Section 3.7]) the identity states that

$$\nabla_\theta \log p(Y;\theta^{(i)}) = \int \nabla_\theta \log p(Y,W;\theta)|_{\theta=\theta^{(i)}} \; p(W|Y;\theta^{(i)})\, \mathrm{d}W. \tag{3.47}$$

It is not difficult to see why this is true; notice that using (3.13), it holds that

$$\begin{aligned}
\nabla_\theta \log p(Y;\theta) &= \frac{1}{p(Y;\theta)} \int \nabla_\theta p(Y,W;\theta)\, \mathrm{d}W \\
&= \frac{1}{p(Y;\theta)} \int \frac{\nabla_\theta p(Y,W;\theta)}{p(W|Y;\theta)} p(W|Y;\theta)\, \mathrm{d}W \\
&= \int \frac{\nabla_\theta p(Y,W;\theta)}{p(W|Y;\theta)p(Y;\theta)} p(W|Y;\theta)\, \mathrm{d}W \\
&= \int \frac{\nabla_\theta p(Y,W;\theta)}{p(Y,W;\theta)} p(W|Y;\theta)\, \mathrm{d}W \\
&= \int \nabla_\theta \log p(Y,W;\theta) p(W|Y;\theta)\, \mathrm{d}W.
\end{aligned} \tag{3.48}$$

This relation can be used to compute an approximate gradient $\nabla_\theta \log p(Y;\theta^{(i)})$ if required. It has not been used in any of the methods of this thesis, however we mention it here for completeness. To get an approximation, we may use the multivariate Gaussian (3.19) obtained by Laplace's method in the place of the posterior in (3.47); however this will give reasonable approximations only if the posterior can be described well with a Gaussian PDF. In this case, we end up with a similar equation as (3.24); thus,

$$\nabla_\theta \log p(Y;\theta^{(i)}) \approx \nabla_\theta Q(\theta,\theta^{(i)})\big|_{\theta=\theta^{(i)}} \tag{3.49}$$

where $Q(\theta, \theta^{(i)})$ is given by (3.24) and we assume that the gradient operator $\nabla_\theta$ commutes with the expectation operator. Unfortunately, we end up with analytically intractable expectations and further approximations are required (using MC sampling for example). Observe that this approximation is exact for linear Gaussian models.

### 3.3.3 Summary

So far, we have presented two main possible approximation methods based on Laplace's approximation method. The results in Example 3.3.3 show that when compared to the true MLE, the sequence $\theta^{(i)}$ of Algorithm 2 does not always converge to a reasonable approximation and no guarantees can be made. Apart from the slow convergence, the behavior of the algorithm deteriorates once the variance of the unobserved process $\boldsymbol{w}_t$ is increased. On the other hand, Examples 3.3.5, 3.3.6 and 3.3.7 indicate that direct optimization of the Laplace's approximation of the log-likelihood function defined in (3.22) might lead to acceptable results. As indicated in the numerical example, the obtained approximations depend on the true model and the used input.

In the following section, we will present possible numerical approximation approaches and discuss their difficulties.

## 3.4 Numerical Approximations

Analytical approximations to some of the quantities in Table 3.1 were considered in the previous part. In this section, we consider possible numerical approximations based on either deterministic numerical integration or Monte Carlo integration (see Appendix A).

### 3.4.1 Approximations Based on Deterministic Numerical Integration – a Case with Independent Outputs

Consider the special case when both $\boldsymbol{y}$ and $\boldsymbol{w}$ are independent and mutually independent stochastic processes, such that we can write

$$p(\boldsymbol{Y}, \boldsymbol{W}; \theta) = \prod_{t=1}^{N} p(\boldsymbol{y}_t, \boldsymbol{w}_t; \theta). \tag{3.50}$$

This is the case for stochastic Wiener models with independent white process disturbance which were used in most of the numerical examples in the previous sections. The likelihood function in this case is

$$p(Y; \theta) = \int p(Y, W; \theta) \, \mathrm{d}W = \int \prod_{t=1}^{N} p(y_t, w_t) \, \mathrm{d}W = \prod_{t=1}^{N} \int p(y_t, w_t) \, \mathrm{d}w_t, \tag{3.51}$$

and instead of having one multidimensional integral over $\mathbb{R}^{d_w N}$, we have $N$ integrals each of dimension $d_w$. Consequently, the likelihood function can be approximated by approximating the integrals

$$p(y_t;\theta) = \int_{\mathbb{R}^{d_w}} p(y_t, w_t;\theta)\,\mathrm{d}w_t = \int_{\mathbb{R}^{d_w}} p(y_t|w_t;\theta)p(w_t;\theta)\,\mathrm{d}w_t$$

which has a known integrand. When the dimension of the unobserved process $\boldsymbol{w}$ is small enough, these integrals can be approximated efficiently using any of the available deterministic numerical integration methods (see [37, Chapter 5]). Observe that the posterior of $\boldsymbol{w}_t$ is never computed in this approach. Approximations of gradient-based algorithms developed on top of deterministic numerical integration were considered in [58]. Here, we will investigate the possibility of employing the EM algorithm.

To proceed, we first make the following two observations. First, note that the posterior can be factorized, due to the independence of the outputs, as

$$p(\boldsymbol{W}|Y;\theta) = \prod_{t=1}^{N} p(\boldsymbol{w}_t|y_t;\theta).$$

This is easy to see if Bayes' theorem is used to write

$$p(\boldsymbol{W}|Y;\theta) = \frac{p(Y,\boldsymbol{W};\theta)}{p(Y;\theta)} = \frac{\prod_{t=1}^{N} p(y_t,\boldsymbol{w}_t)}{\prod_{t=1}^{N} p(y_t;\theta)} = \frac{\prod_{t=1}^{N} p(\boldsymbol{w}_t|y_t;\theta)\prod_{t=1}^{N} p(y_t;\theta)}{\prod_{t=1}^{N} p(y_t;\theta)}$$

Second, we recall from Lemma 3.3.6 that (in general) the intermediate quantity as a function of $\theta$ satisfies

$$Q(\theta,\theta^{(i)}) \propto \int \log p(Y,W;\theta)\;\; p(Y,W;\theta^{(i)})\,\mathrm{d}W. \tag{3.52}$$

In particular, for cases where (3.50) holds, the intermediate quantity satisfies

$$\begin{aligned} Q(\theta,\theta^{(i)}) &\propto \int \log\left(\prod_{t=1}^{N} p(y_t,w_t;\theta)\right)\; \prod_{t=1}^{N} p(y_t,w_t;\theta^{(i)})\,\mathrm{d}W \\ &= \sum_{t=1}^{N}\int \log p(y_t,w_t;\theta)\;\; p(y_t|w_t;\theta^{(i)})p(w_t;\theta^{(i)})\,\mathrm{d}w_t \qquad (3.53) \\ &= \sum_{t=1}^{N}\mathbb{E}\left[p(y_t|\boldsymbol{w}_t;\theta^{(i)})\;\;\log p(y_t,\boldsymbol{w}_t;\theta);\theta^{(i)}\right]. \end{aligned}$$

The maximization of the intermediate quantity of the EM algorithm is then equivalent to the maximization of the right-hand side of (3.53). When the dimension $d_w$ of the process $\boldsymbol{w}$ is small (one or two-dimensional process for example), deterministic numerical integration can be used to approximate the expectations in (3.53) and obtain an approximate ML estimate.

It should be observed that in general, and depending on the parameterization, the maximizer of the $Q$-function will not be available in closed-form. Therefore, the M-step has to be solved numerically using gradient-based algorithms. For this purpose, we need to evaluate the gradient

$$\sum_{t=1}^{N} \int \nabla_\theta \log p(y_t, w_t; \theta) \;\; p(y_t, w_t; \theta^{(i)}) \, \mathrm{d}w_t$$

using deterministic numerical integration. Observe that we assumed that we are allowed to differentiate under the integral sign.

In summary, the EM algorithm is defined by the updates

$$\theta^{(i+1)} = \arg\max_\theta \sum_{t=1}^{N} \int_{\mathbb{R}^{d_w}} \log p(y_t, w_t; \theta) \;\; p(y_t, w_t; \theta^{(i)}) \, \mathrm{d}w_t.$$

Comparing this to a quasi-Newton algorithm when used to evaluate

$$\hat{\theta} := \arg\max_\theta \;\; \sum_{t=1}^{N} \log \;\int_{\mathbb{R}^{d_w}} p(y_t, w_t; \theta) \, \mathrm{d}w_t,$$

we conclude that the EM algorithm has no clear computational advantage for cases where deterministic numerical integration can be used, especially when a faster gradient-based algorithm is applicable. However, we should recall that the iterations of the EM algorithm do not require the global maximizer of the M-step and it seems that the EM algorithm might avoid local maxima which can be problematic for a gradient based algorithm (see [125]). As we discussed earlier it is also preferred to work with the logarithm of the density; however, this is not crucial when $d_w$ is small.

### A limitation

For a general case, with multidimensional integrals of size $d_w N$, integration methods that are based on deterministic griding are hopeless. To see this, let us assume that we have a scalar model with $d_w = 1$ and assume that $N = 100$, a small sample size for system identification problems in most engineering applications. Furthermore, let us assume a very coarse grid of the integral domain of only 10 points per dimension. In this situation, and to get a very rough approximation based on this coarse deterministic grid, we need $10^{100}$ evaluations of the integrand. This is an unimaginable number, much larger than the Eddington number ($N_{\text{Edd}} \approx 10^{79}$, an estimate of the number of protons in the observable universe, see [34]). In such cases, numerical approximations should be based on the probabilistic properties of the integration variable; this is the idea of Monte Carlo methods.

### 3.4.2 Approximations Based on Monte Carlo Simulations

In this part, we will look at approximations of the EM and quasi-Newton algorithms based on the Monte Carlo method. We remind the reader that a brief overview of the Monte Carlo method is given in Appendix A.

A major advantage of Monte Carlo methods is the possibility of analyzing the asymptotic behavior of the approximation and sometimes bounding the approximation errors. In most cases, (theoretical) convergence to the true MLE can be established under mild conditions. Thus, one may consider most of the Monte Carlo methods as exact methods because the approximation errors are functions of only the available computational time and resources. Unfortunately, Monte Carlo methods targeting the MLE can be computationally expensive and its application, to the best of the author's knowledge, is so far limited to problems of small dimensions. We will touch upon this in the coming parts of this chapter before moving to the next chapter where we introduce a PEM that can be used to construct cheaper to computer consistent estimators.

#### The Monte Carlo Expectation-Maximization algorithm

The idea of the Monte Carlo Expectation-Maximization (MCEM) algorithm is due to [144] who considered cases where the mapping $\mathcal{A}$ (of the EM algorithm, see (3.10)) is analytically intractable due to an intractable E-step; however, it assumed that it is easy to obtain random samples according to the (assumed known) posterior distribution.

The principle consists of approximating the E-step using a Monte Carlo sum; we let

$$Q(\theta,\theta^{(i-1)}) \approx Q_{M_i}(\theta,\theta^{(i-1)}) = \frac{1}{M_i}\sum_{m=1}^{N} \log p(Y, W^{(i,m)};\theta)$$

where $\boldsymbol{W}^{(i,1)},\ldots,\boldsymbol{W}^{(i,M_i)}$ are conditionally i.i.d. random variables given the set of random variables

$$\mathcal{F}^{(i-1)} := \{\boldsymbol{\theta}^{(\mathbf{0})}\} \cup \left\{\boldsymbol{W}^{(j,m)} : j = 1,\ldots,i-1,\ m = 1,\ldots,M_j\right\}$$

and distributed according to $p(\boldsymbol{W}|Y;\theta^{(i-1)})$. The first superscript $i$ of $W^{(i,m)}$ is the EM iteration index and the second superscript $m$ is the sample index such that

$$\boldsymbol{W}^{(i,m)} \sim p(\boldsymbol{W}|Y;\theta^{(i-1)}), \quad m = 1,\ldots,M_i, \text{ and } i \in \mathbb{N}.$$

As indicated by the notations, the number of samples $M_i$ can be iteration-dependent. It is also assumed that it is possible to easily realize (simulate) these random variables. The M-step is then replaced by the maximization of the approximate intermediate quantity $Q_{M_{i+1}}(\theta,\theta^{(i)})$. The procedure is summarized in Algorithm 4.

The idea is intuitive and simple, however the implementation of the algorithm and its analysis can get quite involved. Due to the conditional independence assumption,

**Algorithm 4:** The Monte Carlo EM (MCEM) algorithm [144]

```
input  : An initial guess $\theta^{(0)}$, the data $(Y, U)$, ideal sampler of $p(\boldsymbol{W}|Y;\theta)$
         for every feasible $\theta$, and a convergence (stopping) criterion
output : An approximate local maximum of the likelihood function $\hat{\theta}$
Initialize $\theta^{(i)}$ and set index $i \leftarrow 0$
while not converged do
    E-step:  Sample $\boldsymbol{W}^{(i+1,m)}, m = 1, \dots, M_{i+1}$ conditionally
             independently given $\mathcal{F}^{(i)}$ according to $p(\boldsymbol{W}|Y, \theta^{(i)})$,
             then compute
             $Q_{M_{i+1}}(\theta, \theta^{(i)}) = \frac{1}{M_{i+1}} \sum_{m=1}^{N} \log p(Y, W^{(i+1,m)}; \theta^{(i)})$.
    M-step:  Compute $\theta^{(i+1)} \in \arg\max_{\theta \in \Theta} Q_{M_{i+1}}(\theta, \theta^{(i)})$
    $i \leftarrow i + 1$
end
Set $\hat{\theta} = \theta^{(i)}$
```

it is possible under some conditions (see [43] or [102] for example) to prove the convergence of the MCEM algorithm if the number of simulations $M_i$ approaches infinity fast enough as the algorithm approaches convergence (that is as $i \to \infty$). The analysis is based on the observation that each iteration of the MCEM algorithm can be seen as a perturbed version of an EM iteration. The size of the perturbation depends on the history of the Monte Carlo approximation errors and consequently on the sequence $\{M_i\}$. It is important to note that the MCEM is not a monotone algorithm, unlike the EM, and more work has to be done to establish its convergence. The available proofs assume a model within the exponential family; the simplest proof can be found in [20] for example where it is assumed that exact i.i.d. samples are used, and that the joint PDF is, as in (3.14), a member of the exponential family. These assumptions are quite restrictive, especially the independent samples assumption.

For the general models considered in this thesis, the posterior $p(\boldsymbol{W}|Y;\theta)$ is not available and obtaining i.i.d. samples that minimize the Monte Carlo approximation errors is not a trivial task. In principle, the prior $p(\boldsymbol{W};\theta)$ can be used to generate i.i.d. samples which are then weighted with the conditional likelihood function (see (3.40)). However, almost all the weights will be very close to zero and using this (inefficient) method will result in a very high Monte Carlo variance. Approximate methods of sampling that can be used to avoid this difficulty include: Markov Chain Monte Carlo (MCMC) methods, and, depending on the used model, sequential Monte Carlo (SMC) samplers might also be used. Unfortunately, beside the increased computational cost, using approximate sampling methods complicates the algorithm and introduces further errors due to the dependence between the generated samples. Consequently, the analysis of the algorithm gets more involved. See [43] for example for the convergence analysis when $p(\boldsymbol{Y}, \boldsymbol{W}; \theta)$ is a member of the exponential family and when the used samples are based on an MCMC kernel.

The MCEM algorithm based on sequential Monte Carlo (particle filter/smoother) has been used for nonlinear state-space models identification in [106, 109, 125, 146, 147]. Ideas based on sequential Monte Carlo samples are the topic of active research and most of the recent literature on nonlinear system identification relies on it. However, so far, its application is hampered by the two fundamental difficulties: particle degeneracy and impoverishment (see [33, 71]).

In the following part, we would like to check if the methods suggested in Section 3.3 can be improved using the Monte Carlo idea. Nevertheless, it should be kept in mind that approximating high-dimensional distributions by random sampling is a difficult problem and using Monte Carlo sampling in high-dimensional spaces adds to the computational complexity of the methods.

According to Proposition 3.3.4, under Assumption 3.3.3, and by using Laplace's approximation (3.19) the intermediate quantity of the EM algorithm (up to $\theta$-independent terms) is given by (3.24). We also found that two of the three terms that depend on $W$ in the expansion of $p(Y, \boldsymbol{W}; \theta)$ have analytically intractable expectations with respect to (3.19). A natural approximation of these expectations is given by the Monte Carlo estimates

$$\begin{aligned} &\mathbb{E}[\mu^\top(\boldsymbol{W};\theta)\Sigma_Y^{-1}(\theta)\mu(\boldsymbol{W};\theta)|Y;\theta^{(i-1)}] \\ &\qquad \approx \frac{1}{M_i}\sum_{m=1}^{M_i}\mu^\top(W^{(i,m)};\theta)\Sigma_Y^{-1}(\theta)\mu(W^{(i,m)};\theta), \end{aligned} \tag{3.54}$$

and

$$\mathbb{E}[\mu(\boldsymbol{W};\theta)|Y;\theta^{(i-1)}] \approx \frac{1}{M_i}\sum_{m=1}^{M_i}\mu(W^{(i,m)};\theta) \tag{3.55}$$

in which $W^{(i,m)}, m = 1, \ldots, M_i, i = 1, 2, \ldots$ are realizations of the conditionally independent random variables

$$\boldsymbol{W}^{(i,m)} \sim \mathcal{N}\left(\widehat{W}(\theta^{(i-1)}), \Sigma(\widehat{W}(\theta^{(i-1)}), \theta^{(i-1)})\right), \quad m = 1, \ldots, M_i, i = 1, 2, \ldots$$

Using these approximation in (3.24), we compute $Q_{M_i}(\theta, \theta^{(i-1)})$ which is then maximized in the M-step using an iterative numerical optimization algorithm. However, we do not gain much by doing this because Algorithm 2 itself does not have any guarantees. If we let $M_i \to \infty$ as $i$ grows, the sequence $\theta^{(i)}$ will converge to whatever limit Algorithm 2 has. But even for cases where these expectations were tractable, see Example 3.3.3, the algorithm did not perform well. Nevertheless, this problem can be solved by weighting the used samples as explained below.

Observe that the approximations (3.54) and (3.55) are asymptotically exact if the samples are generated according to the true posterior. Therefore, the idea that suggests itself here is the correction of these MC estimators by introducing sample-dependent weights in the MC sums. This idea is known as importance sampling (see Section A.3 in Appendix A).

As pointed out in Lemma 3.3.6, the normalization constant of $p(\boldsymbol{W}|Y;\theta^{(i)})$ is not needed to solve the M-step and the intermediate quantity of the EM algorithm (seen as a function of $\theta$) is in general proportional to the integral

$$\int \log p(Y,W;\theta)\ \ p(Y,W;\theta^{(i)})\,\mathrm{d}W. \tag{3.56}$$

The integrand of this integral is known in terms of a closed-form expression and therefore importance sampling can be used without worrying about weights self-normalization. It can be easily shown (see [117, Theorem 3.12]) that the optimal importance sampling density is given by

$$\tilde{p}_\star(\boldsymbol{W}|Y;\theta,\theta^{(i)}) = \frac{\log p(Y,\boldsymbol{W};\theta)\ \ p(Y,\boldsymbol{W};\theta^{(i)})}{\int \log p(Y,W;\theta)\ \ p(Y,W;\theta^{(i)})\,\mathrm{d}W}$$

which depends on $\theta, \theta^{(i)}$ and $Y$. This PDF is optimal in the sense of minimizing the variance of the approximation error (in fact reducing it to zero). Of course, this optimal density is always unknown as it depends on the intractable integral to be approximated. However, it is possible to use an importance sampling density based on Laplace's approximation method.

Laplace's approximation of the posterior (3.19) is based on the joint PDF $p(\boldsymbol{Y},\boldsymbol{W};\theta)$ and using it as an importance sampling density in (3.56) is not optimal. Note that, even the true posterior itself is not an optimal importance sampling density for the integral in (3.56) (unlike the case when approximating the likelihood function, see the next section). Even in the case of linear models, where Laplace's approximation of the posterior is exact, many samples should be used to ensure small approximation errors. Thus, it is expected that a large number of samples would be needed for non-Gaussian posteriors. Furthermore, to ensure the convergence of the MCEM algorithm, the number of samples should grow with each iteration which makes the method computationally expensive.

### Importance sampling for the E-step

Notice that the integral in Lemma 3.3.6 is alternatively given by

$$\int \frac{\log p(Y,W;\theta)\ \ p(Y,W;\theta^{(i)})}{\tilde{p}(W|Y,\theta,\theta^{(i)})}\tilde{p}(W|Y,\theta,\theta^{(i)})\,\mathrm{d}W$$

which can be seen as an expectation and

$$Q(\theta,\theta^{(i)}) \propto \mathbb{E}\left[\frac{\log p(Y,\boldsymbol{W};\theta)\cdot p(Y,\boldsymbol{W};\theta^{(i)})}{\tilde{p}(\boldsymbol{W}|Y,\theta,\theta^{(i)})};\theta,\theta^{(i)}\right]$$

in which $\mathbb{E}$ is with respect to an importance sampling density $\tilde{p}(\boldsymbol{W}|Y,\theta,\theta^{(i)})$. We will assume that (3.19) is used as an importance sampling density. A Monte Carlo approximation of the E-step (up to $\theta$-independent constant) at iteration $i$ is then given by

$$\frac{1}{M_{i+1}}\sum_{m=1}^{M_{i+1}}\frac{\log p(Y,W^{(i,m)}(\theta);\theta)\cdot p(Y,W^{(i,m)}(\theta);\theta^{(i)})}{\tilde{p}(W^{(i,m)}(\theta)|Y,\theta)}$$

in which $W^{(i,m)}(\theta)$ are i.i.d. samples according to (3.19). We summarize the method in Algorithm 5.

**Algorithm 5:** The Monte Carlo EM (MCEM) algorithm based on importance sampling

```
input  : An initial guess $\theta^{(0)}$, the data $(Y,U)$, and a convergence
         (stopping) criterion
output : An approximate local maximum of the likelihood function $\hat{\theta}$
Initialize $\theta^{(i)}$ and set index $i \leftarrow 0$
while not converged do
    E-step:  Find an importance sampling PDF $\tilde{p}$ (by solving (3.17)
             or by any alternative method) and simulate
             $\boldsymbol{W}^{(i+1,m)} \sim \tilde{p},\ m = 1,\dots,M_{i+1}$.
    M-step:  $\theta^{(i+1)} \in \arg\max_{\theta} \sum_{m=1}^{M_{i+1}} \frac{\log p(Y,W^{(i+1,m)};\theta)\cdot p(Y,W^{(i+1,m)};\theta^{(i)})}{M_{i+1}\,\tilde{p}(W^{(i+1,m)}|Y,\theta,\theta^{(i)})}$
    $i \leftarrow i+1$
end
Set $\hat{\theta} = \theta^{(i)}$
```

**Remark 3.4.1** (MCEM based on importance sampling)**.**

- *The generality of Algorithm 5 hides the difficulty of the original problem in the step of choosing an importance sampling density. A careless choice will result in very small importance weights that are practically* 0*. In this case, most or even all the samples will not contribute to the approximation of the integral.*

- *An advantage of the use of importance sampling, as described in Algorithm 5, is the possibility of using common random numbers (see Section A.2). This preserves the continuity properties of the function to be maximized in the M-step (see [49]). In addition, generating the required i.i.d. samples becomes an easy task (e.g., simulating Gaussian random variables).*

- *In contrast to estimators based on MCMC/SMC samplers, the Monte Carlo estimator used in the E-step is an unbiased estimator of (3.52) for any finite value $M_i$. The reason is that self-normalization of the weights is not required.*

- *To reduce the computational demand of the algorithm, the same importance sampling density can be used for a few iterations before recomputing a new* *density. However, this would require even larger number of samples to control the possible increase in the approximation error variance.*

The difficulties of applying the above method are illustrated in the next numerical example.

**Example 3.4.1** (Expectation-Maximization based on Laplace importance sampling)**.** We first demonstrate the performance of Algorithm 5 on the simple linear state-space model

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \theta \boldsymbol{x}_t + u_t + \boldsymbol{w}_t, \qquad && \theta = 0.7, \quad x_0 = 0, \\ \boldsymbol{y}_t &= \boldsymbol{x}_t + \boldsymbol{v}_t, && t = 1, \dots, N = 100. \end{aligned} \tag{3.57}$$

For each $t$, the input $u_t$ is a known realization of a Gaussian random variable with mean 1 and variance 1, and $\boldsymbol{w}_t, \boldsymbol{v}_t \sim \mathcal{N}(0,1)$ such that they are mutually independent. We used $M_i = 1000$ for all $i$, and terminated the algorithm once $\|\theta^{(i)} - \theta^{(i-1)}\| < 10^{-5}$. The used importance sampling density is given by (3.19) which coincides in this case (see Example 3.3.2) with the true posterior. Therefore all the values $p(Y, \boldsymbol{W}^{(i,m)}; \theta^{(i)})/\tilde{p}(\boldsymbol{W}^{(i,m)}|Y, \theta^{(i)})$ are the same regardless of $m$ and are equal to $p(Y; \theta^{(i)})$. The results of one realization are shown in Figure 3.14; we see that the iterations of the algorithm converge to the true MLE which agrees with the above theoretical motivation.

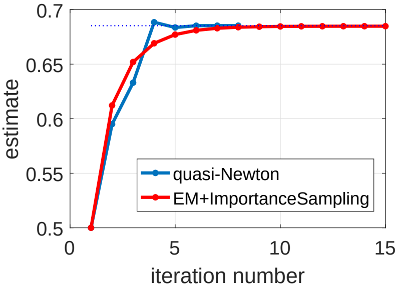


**Figure 3.14:** The MLE of $\theta$ in (3.57). The blue curve shows the iterates of a quasi-Newton algorithm optimizing the true likelihood, and the red curve shows the iterations of Algorithm 5 in which Laplace's approximation (3.19) was used as a sampling density.

We now consider the nonlinear model (3.37) from Example 3.3.3. In order to demonstrate the behavior of Algorithm 5 in comparison to Algorithm 2 we assumed that $N = 5$. This is a very short sample size for many practical applications, and is used only to avoid several numerical issues (see below) that are present when running the algorithm for larger $N$. Furthermore, Laplace's approximation of the posterior, given by (3.19), is used as an importance sampling density.

Figure 3.15 shows the simulation results of both algorithms when $\lambda_w = 1$ (a case which was difficult for Algorithm 2). Apart from the slow convergence, the results show that the MCEM with importance sampling converges to the true ML estimate as expected, unlike the EM algorithm based on Laplace's approximation for the posterior. It is also clear that the use of common random numbers preserves the continuity and smoothness of the intermediate value function. These results highlight the advantage when $N$ is small.

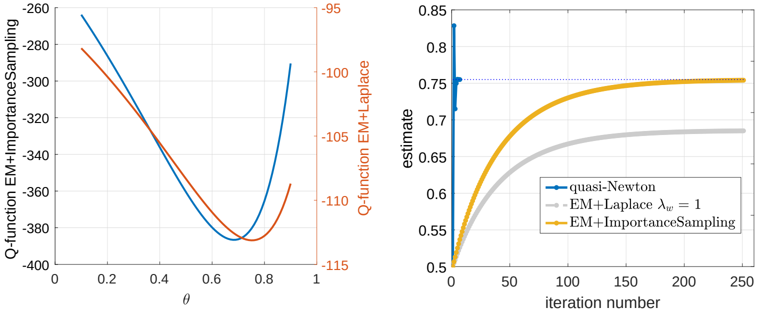


**Figure 3.15:** The plot shows a comparison between Algorithm 2 (in grey) and Algorithm 5 (in yellow) for $N = 5$ and $M_i = 10^4$ for all $i$. The plots on the left show the $Q$-function at iteration 200. On the right, the iterations of the quasi-Newton algorithm optimizing the true likelihood are shown against the iterations of the two EM algorithms.

Unfortunately, the situation gets more complicated when $N$ increases. The simulation results for one realization when $N = 100$, $M_i = 10^4$ for all $i$ are shown in Figure 3.16. We see that Algorithm 5 does not converge to the true ML estimate. The reason for this becomes clear when we look at the weights, normalized by dividing by their maximum such that all the weights are in between 0 and 1. For example at iteration 200, among $10^4$ samples of the 100-dimensional vector $\boldsymbol{W}$, only 28 samples come with a weight larger than 0.2 and almost all the rest have weights very close to zero. This indicates that the importance sampling density obtained by Laplace's approximation method does not capture the important regions of the support of the posterior of $\boldsymbol{W}$ and is not the optimal choice.

---

The previous example demonstrated possible advantages and disadvantages of Algorithm 5. One major advantage is the use of conditionally i.i.d. samples that makes the convergence of the algorithm to the true MLE possible if $M_i \to \infty$ fast enough. The major difficulty lies in the choice of the proposal density and the computations of the weights. Inefficient densities (i.e., those that are not close in shape to the optimal sampling density) will lead to very small weights for almost all the samples. Another numerical difficulty is how the weights are computed; observe that the weights are given by the ratio of high-dimensional PDFs, and suitable

[6] Normalized by dividing by the largest weight, such that all the weights are in $[0, 1]$.

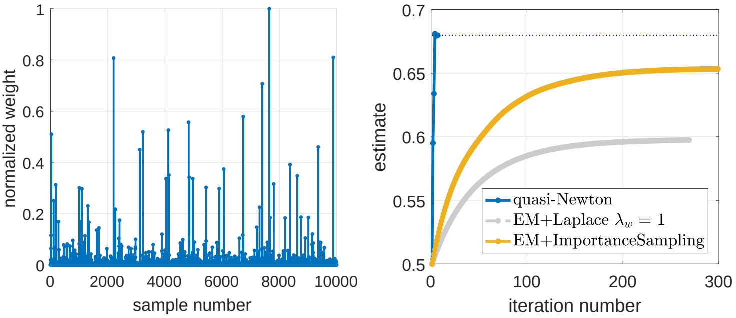


**Figure 3.16:** The plot shows a comparison between Algorithm 2 (in grey and red) and Algorithm 5 (in yellow and blue) for $N = 100$ and $M_i = 10^4$ for all $i$. The plot on the left shows the normalized weights[6] used in Algorithm 5 at iteration 200. On the right, the iterations of the quasi-Newton algorithm optimizing the true likelihood is shown against the iterations of the two EM algorithms.

$\theta$-independent normalization constants have to be used for each $i$ to make the values tractable.

For sequential models, like state-space models, these difficulties can be alleviated by the use of sequential Monte Carlo samplers as suggested in [109, 125] for example. The idea there is to propagate the samples of $\boldsymbol{w}_t$ sequentially over $t$ such that the weights are computed and normalized in the low dimension $d_w$. While such methods are able to keep the weights relatively high, the required number of samples remains large and the methods are so far applicable only to problems with small dimensions.

In general, Markov Chain Monte Carlo methods can be used to sample from the posterior. However, once more, the difficulty of the problem is transferred to the choice of the proposal density of the MCMC kernel. In addition, a large number of samples might be required to ensure the convergence of the Markov Chain to the stationary distribution.

One of the problems of Algorithm 5 is the difficulty of finding the optimal importance sampling density. As we pointed out above, the true posterior is not optimal and even in the linear case $M_i$ should be large enough to reduce the MC approximation errors. The situation might be slightly different if we try instead to use the importance sampling idea to approximate the likelihood function. As shown in the next part, the true posterior is the optimal importance sampling density when used to approximate the marginalization integral defining the likelihood function.

### Monte Carlo approximations of gradient-based methods

Monte Carlo approximations can be alternatively used to approximate the likelihood function itself. The idea is then to use the approximation within a quasi-Newton algorithm where the gradients are estimated numerically using finite differences. For this approach to work we need the estimators to use common random numbers.

Recall that the likelihood function is given by the marginalization integral

$$p(Y;\theta) = \int p(Y,W;\theta)\,\mathrm{d}W = \int p(Y|W;\theta)p(W;\theta)\,\mathrm{d}W.$$

In general, an unbiased Monte Carlo estimator of the likelihood function at a given value $\theta$ and observation vector $Y$ can be obtained using i.i.d. samples

$$\boldsymbol{W}^{(m)}(\theta) \sim p(\boldsymbol{W};\theta), \quad m = 1,\dots,M,$$

which are used to write

$$\widehat{p(Y;\theta)} = \frac{1}{M}\sum_{m=1}^{M} p(Y|W^{(m)}(\theta);\theta). \tag{3.58}$$

It is trivial to see that

$$\mathbb{E}[\widehat{p(Y;\theta)}] = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}[\widehat{p(Y|\boldsymbol{W}^{(m)}\theta)}] = p(Y;\theta),$$

and under the assumption that $p(Y|\boldsymbol{W}^{(m)}(\theta);\theta)$ has a finite variance with respect to $p(W;\theta)\,\mathrm{d}W$, it also holds that

$$\mathbf{var}\left[\widehat{p(Y;\theta)};\theta\right] = \frac{1}{M}\,\mathbf{var}[p(Y|\boldsymbol{W}^{(m)}(\theta);\theta);\theta].$$

A notable feature of this estimator is that the variance does not depend directly on the dimension of either $W$ or $Y$ and it holds in general that the approximation error is $\mathcal{O}(M^{-1})$. This is an advantage of Monte Carlo approximation methods over deterministic methods. Furthermore, due to the use of independent samples, a direct application of the strong law of large numbers shows that

$$\frac{1}{M}\sum_{m=1}^{M} p(Y|W^{(m)}(\theta);\theta) \xrightarrow{\text{a.s.}} p(Y;\theta) \quad \text{as} \quad M \to \infty$$

and the precise rate of convergence is given by the law of the iterated logarithm (see [23, Section 7.6]). Moreover, the standard version of the central limit theorem of i.i.d. random variables implies that

$$\frac{\widehat{p(Y;\theta)} - \mathbb{E}[p(Y|\boldsymbol{W};\theta);\theta]}{\widehat{\mathbf{var}}\left[\widehat{p(Y;\theta)};\theta\right]} \text{ is approximately } \mathcal{N}(0,1)$$

where

$$\widehat{\mathbf{var}}[\widehat{p(Y;\theta)};\theta] = \frac{1}{M^2}\sum_{m=1}^{M}[p(Y|W^{(m)}(\theta)) - \widehat{p(Y;\theta)}]^2.$$

The knowledge of the asymptotic distribution can be used for the construction of confidence intervals and various tests for the convergence of the estimator.

These are all desirable "asymptotic" properties; however, the accuracy of (3.58) depends significantly on the nature of the conditional likelihood and it is usually the case that $\mathbf{var}[p(Y|\boldsymbol{W}^{(m)}(\theta);\theta);\theta]$ is very large for large $d_w N$. Consequently, prohibitively large $M$ might be needed to arrive at a reasonable approximation. Observe that the samples $\boldsymbol{W}^{(m)}$ do not rely on the specific realization $Y$; they only rely on the prior PDF which is usually not concentrated. Also note that for a fixed $Y$ and $\theta$, the function $p(Y|W;\theta)$ seen as a function of $W$ has most of its mass in regions whose volume is only a fraction of the total mass of the prior. To give a hint on how this situation looks like, we give the following example.

**Example 3.4.2.** Consider the following static linear model

$$\boldsymbol{y} = \theta\boldsymbol{w} + \boldsymbol{v}$$

with $\theta = 50$ and both $\boldsymbol{w}$ and $\boldsymbol{v}$ are independent random variables such that $\boldsymbol{w} \sim \mathcal{N}(0,1)$ and $\boldsymbol{v} \sim \mathcal{N}(0,0.1)$. Figure 3.17 shows the graph of the function

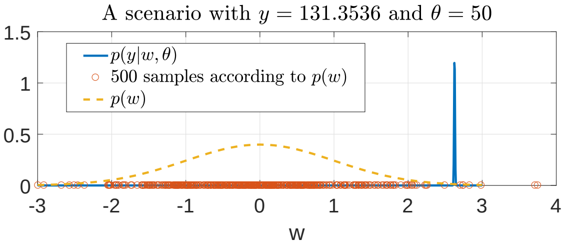


**Figure 3.17:** Illustration of the inefficiency of prior samplers for integrating a conditional densities

$$p(y|w;\theta) = \frac{1}{\sqrt{2\pi}\sqrt{0.1}}\exp\left(-\frac{1}{2\cdot 0.1}(y-50w)^2\right)$$

when $y = 131.353$ and $w$ is varied. If we consider $w$ as a parameter, this function is, by definition, the likelihood function of $w$. It is clear that even in this one dimensional example, the function $p(y|\cdot;\theta)$ has almost all of its mass, i.e., $\{w : p(y|w\ \theta) > \epsilon\}$ with small $\epsilon > 0$, concentrated in a very short interval; for the given realization, the interval length is around 0.06 for $\epsilon = 10^{-4}$. Among 500 independent samples generated according to the prior $p(\boldsymbol{w})$, none have hit the important interval of the function $p(y|\cdot;\theta)$ to be integrated.

Of course, for the above simple example, increasing the number of samples $M$ will eventually give an acceptable approximation of $p(y;\theta)$. However, when the dimension of $Y$ is large, this problem is magnified and will always be present in practice regardless of what the realizations of the involved variables are or how big $M$ is.

The next example clarifies the inefficiency of the MC estimator (3.58) when used within a quasi-Newton algorithm to approximate the ML estimate. Again, we estimate the gradients with finite differences.

**Example 3.4.3.** To demonstrate the practical inefficiency of using prior samples, consider a simple scalar linear state-space model.

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \theta \boldsymbol{x}_k + \boldsymbol{w}_k, \quad \theta = 0.7, \quad x_0 = 0, \\ \boldsymbol{y}_k &= \quad \boldsymbol{x}_k + e_k, \quad\ \ t = 1, \dots, N. \end{aligned}$$

We assume that $\boldsymbol{w}_t \sim \mathcal{N}(0, 1.5)$ and is independent of $\boldsymbol{v}_t \sim \mathcal{N}(0, 1)$. We let $N = 200$, and simulate the estimator over 1000 realization of $Y$ for a grid of values for $M$ between $10^3$ and $10^4$. The result is shown in comparison to the MLE in Figure 3.18. It is clear that the approximation error, which is reflected in the MSE, is quite large and it seems that $M$ has to be very large to get any acceptable approximations.

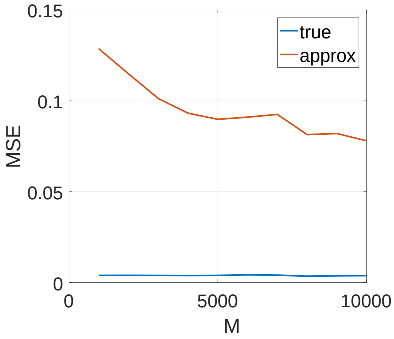


**Figure 3.18:** Illustration of the inefficiency of direct sampling according to the prior: the plot shows the empirical MSE of the estimates as a function of $M$; the blue curve is the MSE of the true ML estimate which is independent of $M$, and the red curve is the MSE of the approximation.

The problem observed in the above example is a manifestation of the curse of dimensionality. Observe that the likelihood function can be bounded in terms of a volume of a set under the measure $p(W;\theta)\,\mathrm{d}W$ since

$$p(Y;\theta) \leq C \int \mathbf{1}_{\mathcal{S}_{Y,\theta}}\, p(W;\theta)\,\mathrm{d}W$$

for some positive constant $C$ and

$$\mathcal{S}_{Y,\theta} := \{W : p(Y|W;\theta) \neq 0\}.$$

Pick a small positive $\epsilon$ and consider the set

$$\tilde{\mathcal{S}} := \{W : p(Y|W;\theta) \geq \epsilon\} \subset \mathcal{S}_{Y,\theta}$$

and observe that the difference (in volume) between these two sets can be made arbitrarily small regardless of $N$ by picking $\epsilon$ small enough. Now notice that the volume of $\tilde{\mathcal{S}}$ under $\mathrm{d}W$ must shrink with the dimension $N$. To see this, let us consider an example of a scalar model for which

$$p(Y|W;\theta) \propto \exp\left(-\frac{1}{2}\|Y - W\|_2^2\right).$$

We then center the coordinates at $Y$ so that

$$\tilde{\mathcal{S}} = \mathcal{B}_N(\tilde{\epsilon}) = \{W : \|W\|_2 \leq \tilde{\epsilon}\}$$

This is a ball of radius $\tilde{\epsilon}$ in $\mathbb{R}^N$ whose volume approach zero as $N$ grows to $\infty$ (see [47, Chapter 1]). This means that $\mathcal{S}_{Y,\theta}$ will have a shrinking volume; thus, for a small fixed value $\epsilon$, the probability that a sample $W$ drawn according to the prior is in $\tilde{\mathcal{S}}$ approaches zero as N increases.

### Importance sampling for the quasi-Newton algorithm

To reduce the variance of the estimates, importance sampling can be used. For every PDF $\tilde{p}(\boldsymbol{W};Y,\theta)$ that may depend on $Y$ and $\theta$, we have the following alternative representation of (3.3)

$$p(Y;\theta) = \int \frac{p(Y|W;\theta)p(W)}{\tilde{p}(W;Y,\theta)}\tilde{p}(W;Y,\theta)\,\mathrm{d}W.$$

An unbiased Monte Carlo estimate based on importance sampling can be obtained using i.i.d. samples

$$\boldsymbol{W}^{(m)}(\theta) \sim \tilde{p}(\boldsymbol{W};Y,\theta), \quad m = 1,\ldots,M,$$

which is used to define the estimate

$$\widehat{p(Y;\theta)} = \frac{1}{M}\sum_{m=1}^{M}\frac{p(Y|W^{(m)}(\theta);\theta)p(W^{(m)}(\theta))}{\tilde{p}(W^{(m)}(\theta);Y,\theta)}. \tag{3.59}$$

Under similar conditions as before,

$$\mathbf{var}\left[\widehat{p(Y;\theta)};\theta\right] = \frac{1}{M}\mathbf{var}\left[\frac{p(Y|\boldsymbol{W}^{(m)}(\theta);\theta)p(\boldsymbol{W}^{(m)}(\theta))}{\tilde{p}(\boldsymbol{W}^{(m)}(\theta);Y,\theta)};\theta\right] \tag{3.60}$$

where the variance operator is with respect to $\tilde{p}(\boldsymbol{W};Y,\theta)$. It is clear that one necessary condition for the variance to be finite is that the ratio

$$\frac{p(W^{(m)}(\theta))}{\tilde{p}(W^{(m)}(\theta);Y,\theta)}$$

has to be bounded.

Now, observe that this variance can be made equal to zero by choosing

$$\tilde{p}(\boldsymbol{W};Y,\theta)=p(\boldsymbol{W}|Y,\theta).$$

Indeed, in this case the variance on the right-hand side of (3.60) becomes

$$\mathbf{var}\left[\frac{p(Y|\boldsymbol{W}^{(m)}(\theta);\theta)p(\boldsymbol{W}^{(m)}(\theta))}{\frac{p(Y|\boldsymbol{W}^{(m)}(\theta);\theta)p(\boldsymbol{W}^{(m)}(\theta))}{p(Y;\theta)}};\theta\right]=\mathbf{var}\left[p(Y;\theta)\right]=0$$

and only one sample is required, i.e., $M=1$. The approximation is then given by

$$\frac{p(Y|W^{(1)};\theta)p(W^{(1)})}{p(W^{(1)}|Y;\theta)}$$

which is independent of the used sample $W^{(1)}$ and is equal to $p(Y;\theta)$. Therefore, we conclude that the optimal sampling density for (3.59) is the posterior of $\boldsymbol{W}$.

An approximation of the MLE can be obtained by using Laplace's approximation of the posterior as an importance sampling density, and then using the estimates of the likelihood function within a quasi-Newton algorithm. In this case, the resulting algorithm is exact for linear Gaussian models. Furthermore, for cases where the posterior has one dominant mode, Laplace's approximation is expected to capture the important region of the posterior support, and therefore the variance of the likelihood estimate will be reduced. We summarize the suggested method in Algorithm 6.

The method described above has several disadvantages. As we pointed out on page 50, from a numerical point of view it is preferred to work with the log-likelihood function. To be able to use Algorithm 6, the value of the likelihood function has to be scaled with an appropriate $\theta$-independent factor.

Assume that both $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are independent zero mean Gaussian random variables with variances $\lambda_w I$ and $\lambda_v I$ respectively. For given $Y$, $\theta$, importance sampling density, and corresponding samples $W^{(m)}$, the likelihood approximation is evaluated by computing the fraction

$$\begin{aligned}&\frac{p(Y|W^{(m)};\theta)p_W(W^{(m)})}{\tilde{p}_W(W^{(m)}|Y;\theta)}=\\&\qquad\frac{c_1\exp\left(-\frac{1}{2\lambda_e}\|Y-\mathcal{M}(U,W^{(m)};\theta)\|^2\right)c_2\exp\left(-\frac{1}{2\lambda_w}\|W^{(m)}\|^2\right)}{c_3(\theta)\exp\left(-\frac{1}{2}(W^{(m)}-\widehat{W}(\theta))^\intercal[\Sigma(\widehat{W};\theta)]^{-1}(W^{(m)}-\widehat{W}(\theta))\right)},\end{aligned}\tag{3.61}$$

with $c_1=\dfrac{1}{(2\pi\lambda_v)^{\frac{N}{2}}}$, $c_1=\dfrac{1}{(2\pi\lambda_w)^{\frac{N}{2}}}$, and $c_3(\theta)=\dfrac{1}{(2\pi)^{\frac{N}{2}}[\det\Sigma(\widehat{W},\theta)]^{\frac{1}{2}}}$.

For large values of $N$, a direct calculation of this expression is not possible. First, the value $[\det\Sigma(\widehat{W},\theta)]^{-\frac{1}{2}}$ will be a very small number. Second, the constants $c_1$ and $c_2$ will be very large. Third, it is likely that the arguments of the exponential function

**Algorithm 6:** Monte Carlo approximation of the quasi-Newton algorithm based on Laplace importance sampling

```
input  : An initial guess $\theta^{(0)}$, the data $(Y, U)$, and a convergence
         (stopping) criterion
output : An approximate local maximum of the likelihood function $\hat{\theta}$
Initialize $\theta^{(i)}$ and set index $i \leftarrow 0$
while not converged do
    Find $\tilde{p}$:   Find an importance sampling PDF $\tilde{p}$ (by solving (3.17)
                 for example, or by any alternative way) and simulate
                 $\boldsymbol{W}^{(i+1,m)} \sim \tilde{p}$, $m = 1, \ldots, M_{i+1}$.
    Update $\theta$:   $\theta^{(i+1)} = \theta^{(i)} - \alpha_i H_i \widehat{\nabla_\theta p(Y;\theta^{(i)})}$
                 where the estimate $\widehat{\nabla_\theta p(Y;\theta^{(i)})}$ is computed using
                 finite differences based on the estimate
                 $\widehat{p(Y;\theta)} = M_{i+1}^{-1} \sum_{m=1}^{M_{i+1}} \frac{p(Y|W^{(i+1,m)};\theta)p(W^{(i+1,m)};\theta)}{\tilde{p}(W^{(i+1,m)};Y,\theta)}$
                 The step $\alpha_i$ is determined by an inexact line search
                 based on $\widehat{p(Y;\theta)}$, and $H_i$ approximates the inverse of
                 the Hessian using the BFGS method for example.
    $i \leftarrow i + 1$
end
Set $\hat{\theta} = \theta^{(i)}$
```

will be too large, making the exponential function equal to zero for any computer with finite precision. Nevertheless, by first taking the logarithm of the fraction in (3.61) and then applying the exponential function to the result the problem can be alleviated. The logarithm transforms products into sums and exponents into scaling factors, which makes the numbers more tractable. In addition, the whole expression can be normalized by any constant that is independent of $\theta$. This can be calculated based on $\theta^{(1)}$ such that the minimum value of the fraction (over $m$) is equal to 1 and then keep it fixed for all future iterations.

Notice that, in principle, an approximation of the gradient of the log-likelihood function can be computed based on importance sampling. According to (3.48)

$$\nabla_\theta \log p(Y;\theta) = \int \nabla_\theta \log p(Y, W;\theta) p(W|Y;\theta) \, \mathrm{d}W$$

Because the posterior $p(W|Y;\theta)$ is only known up to the normalizing constant, self-normalized importance sampling has to be used and the resulting estimator

$$\widehat{\nabla_\theta \log p(Y;\theta)} = \frac{1}{M} \sum_{m=1}^{M} \omega^{(m)} \nabla_\theta \log p(Y, W^{(m)};\theta)$$

where

$$\omega^{(m)} = \frac{\frac{p(Y,W^{(m)};\theta)}{\tilde{p}(W^{(m)};Y,\theta)}}{\sum_{m=1}^{M} \frac{p(Y,W^{(m)};\theta)}{\tilde{p}(W^{(m)};Y,\theta)}}$$

is biased for every finite $M$. This approximation can be used in Algorithm 6 instead of $\widehat{\nabla_\theta p(Y;\theta)}$, but the performance depends significantly on the choice of the importance sampling density. Observe that the true posterior is not the optimal importance sampling density for directly approximating the gradient of the likelihood function. The number of samples $M_i$ has to be large enough to reduce the possible bias and guarantee that the effective sample size is large enough.

**Example 3.4.4.** Consider the following FIR model with cubic nonlinearity at the output

$$\begin{aligned} \boldsymbol{x}_t &= \theta u_t + u_{t-1} + \boldsymbol{w}_t, & \theta &= 0.5 \\ \boldsymbol{y}_t &= \ \ \boldsymbol{x}_t^3 + \boldsymbol{v}_t, & u_0 &= 0,\ t = 1, \dots, N. \end{aligned}$$

in which, $u_t \sim \mathcal{N}(0, \frac{1}{3}), \boldsymbol{v}_t \sim \mathcal{N}(0, 0.1)$, and $\boldsymbol{w}_t \sim \mathcal{N}(0, 0.2)$ are independent. We fixed the number of observations $N = 1000$, and the number of Monte Carlo samples $M = 5000$ and implemented Algorithm 6. Because the outputs are independent over time, we may use numerical integration to accurately approximate the true MLE. Observe that Algorithm 6 does not use the assumption that the outputs are independent over time.

The average results over 1000 disturbance, noise and input realizations are summarized in Table 3.2. The simulation results indicate that the Algorithm 6 is unbiased; but has a higher variance compared to the true MLE due to the Monte Carlo approximation errors.

**Table 3.2:** Simulation results for Example 3.4.4.

| | Mean | std | MSE |
|---|---|---|---|
| Deterministic numerical Integration (true MLE): | 0.4944 | 0.0319 | 0.0010 |
| Algorithm 6: | 0.5151 | 0.0469 | 0.0024 |

### 3.4.3 Summary

We saw that when the output $\boldsymbol{y}$ is an independent process, deterministic numerical integration can be used to approximate the M-step of the EM algorithm whenever the dimension $d_w$ is small enough. We then introduced the Monte Carlo EM (MCEM) algorithm and suggested the use of Laplace's approximation of the posterior $p(\boldsymbol{W}|Y;\theta)$ as an importance sampling density. The resulting algorithm (Algorithm 5) has desirable theoretical guarantees; however as Example 3.4.1 shows, when $N$

increases the variance of the importance sampling weights can be very high due to a poor importance sampling density.

In Algorithm 6 we proposed the use of importance sampling to directly approximate the likelihood function. One main difference between the two algorithms is that in the latter the used importance sampling density is an approximation of the optimal importance sampling density. The implementation of both algorithms has a numerical difficulty due to the use of probabilities of high dimensional vectors when computing the weight.

## 3.5 Conclusions

In this chapter, we studied several approximation approaches to the ML problem of parametric nonlinear models. We focused on two main algorithms: (i) the Expectation-Maximization algorithm, and (ii) the quasi-Newton algorithm and discussed their properties, advantages and disadvantages in Section 3.2. In Section 3.3, we introduced Laplace's approximation method that can be used to obtain a Gaussian approximation of the posterior of $\boldsymbol{W}$ as well an approximation of the likelihood function at a given $\theta$. Several simulation examples show that direct approximation of the likelihood function may be acceptable; the accuracy depends not only on the true model but also on the used input signal.

One disadvantage common to all methods based on analytic (functional) approximations is the difficulty of analyzing the resulting estimates. The best that could be done is to show that these approximations are exact for the linear case. This comes in contrast to approximations based on Monte Carlo methods, as shown in Section 3.4.

Whenever Monte Carlo approximations are used, it is usually possible to establish asymptotic results and in some cases get finite (approximate) sample bounds. Yet, Monte Carlo methods come with their own computational and numerical difficulties. The main challenge is to reduce the number of the required MC samples to a reasonable value and in the same time guarantee that the MC variance is small enough. The current state-of-the-art sampling methods rely on particle filters and (particle) MCMC algorithms (see for example [126]).

Chapter 4

# Linear Prediction Error Methods

The methods of Chapter 3 attempt several approximations of the MLE. Because they rely on approximations of rather complicated PDFs in high-dimensional spaces, they are computationally expensive and come with several numerical difficulties as well as unfavourable properties. In this chapter, we look at the problem from a prediction error perspective where we do not necessarily seek an optimal predictor. We show that it is possible to use computationally attractive suboptimal predictors to construct consistent estimators in a prediction error framework.

## 4.1 Introduction

A fundamental step of any prediction error method is the computation of a predictor for the assumed model. The optimal one-step ahead predictor(see (2.35) on page 34), which minimizes the variance of the prediction errors, is usually the preferred choice. However, as shown in Chapter 1, the optimal predictors of stochastic nonlinear models are, in general, analytically intractable. Approximations of the optimal predictor are as complicated as approximations of the efficient MLE, since both rely on the intractable likelihood function. Fortunately, as discussed in Section 2.4.2, the predictors in the PEMs can be defined in many ways that might even include some ad hoc non-probabilistic arguments (also see [92, Section 3.3]). In this chapter, we are interested in consistent instances of the PEMs based on linear predictors. The obtained results can be seen as extensions of the linear case (see Examples 2.4.2, 2.4.3, and 2.4.4 in Chapter 2), and can be motivated by Wold's decomposition as given in Theorem 2.1.6. We will work under the following assumption:

**Assumption 4.1.1.** *The outputs can be described by a known vector relation* $\mathcal{M}$*,*

$$\boldsymbol{Y} = \mathcal{M}(U, \boldsymbol{W}; \theta) + \boldsymbol{V}, \tag{4.1}$$

*where* $U \in \mathbb{R}^{d_u N}$ *is a vector of known inputs,* $\boldsymbol{W} \in \mathsf{L}_2^{d_w N}$ *is an unobserved random vector with a known (possibly parameterized) PDF, and* $\boldsymbol{V} \in \mathsf{L}_2^{d_y N}$ *is an unobserved random vector representing measurement noise.*

Notice, however, that the methods in this chapter can be easily extended to a more general class of models where $\boldsymbol{V}$ is not necessarily additive. Assume for the moment that $\boldsymbol{v}$ is a linearly filtered white noise with zero mean and known covariance function, but unknown PDFs. Under the assumption that the centered process $\boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t;\theta]$ is purely non-deterministic[1] with zero initial conditions, Wold's decomposition as given in Theorem 2.1.6 ensures the existence of the representation

$$\boldsymbol{Y} = \mu(U,\theta) + \boldsymbol{Z}(U,\theta)$$

in which $\mu(U,\theta)$ is the mean vector of $\boldsymbol{Y}$: namely, $[\mu(U;\theta)]_t := \mathbb{E}[\boldsymbol{y}_t;\theta]$, and

$$\boldsymbol{Z}(U,\theta) = L(U;\theta)\boldsymbol{\mathcal{E}}$$

in which $L(U;\theta)$ is a lower unitriangular matrix of decaying elements, and $\boldsymbol{\mathcal{E}}$ is the vector of innovations; that is, a vector whose elements are white noise[2] with finite covariance such that $[\boldsymbol{\mathcal{E}}]_t \in \boldsymbol{\mathcal{H}}_t$. We invite the reader to compare this representation to the models in (2.21) and (2.27) on pages 26 and 27 respectively.

We see that regardless of the underlying model, as long as the conditions of Theorem 2.1.6 are satisfied, the output vector can always be written in terms of the mean vector plus a causally filtered white noise. It is important to notice that, in general, the filtering matrix $L$ (the equivalent of the noise model) depends on the used input. Furthermore, we observe that Wold's decomposition does not specify the distribution of $\boldsymbol{\mathcal{E}}$ but only its second moments; therefore, it is an incomplete representation characterizing only the first and second moments of the process. Fortunately, this characterization is sufficient for the construction of optimal linear predictors (see Appendix B).

## 4.2 Using Linear Predictors and PEMs for Nonlinear Models

It is important to understand that what we are suggesting here are linear predictors for a process with a nonlinear underlying model. The objective is to estimate the parameters of the assumed nonlinear structure by using a linear predictor in a PEM (a different objective compared to [36]). To do so, we are required to construct a predictor that is parameterized by $\theta$ and is linear in the observations $Y$. However, the dependence on the known inputs $U$ can be nonlinear.

We clarify this in the following example.

[1] i.e., the linear deterministic part of $\boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t;\theta]$ is zero.

[2] Recall that we define white noise as a sequence of uncorrelated random variables but not necessarily independent, and that $\boldsymbol{\mathcal{H}}_t := \overline{\mathbf{sp}}\{\boldsymbol{y}_s : s \leq t\}\ \forall t \in \mathbb{Z}$.

**Example 4.2.1** (Linear prediction of a non-stationary process)**.** Consider the second-order discrete-time stochastic process

$$\boldsymbol{y}_t = \mu_t(\theta) + \boldsymbol{\zeta}_t, \quad t \in \mathbb{Z},$$

and assume for the moment that $\boldsymbol{\zeta}$ is a purely non-deterministic zero mean stationary process with strictly positive rational power spectrum. The deterministic signal $\mu_t(\theta)$ is assumed to be generated according to some given recursion; for example,

$$\mu_t(\theta) = f(\mu_{t-1}(\theta), \ldots, \mu_{t-n_\mu}(\theta), u_{t-1}, \ldots, u_{t-n_u}; \theta), \quad n_\mu, n_u \in \mathbb{N}$$

for some known parameterized function $f$ and known deterministic sequence $\{u_k\}$, but it can be generated in any other way. Observe that we did not specify the distribution of $\boldsymbol{\zeta}$ and that $\boldsymbol{y}$ is given in the form of Wold's decomposition. Indeed, the signal $\mu_t(\theta) = \mathbb{E}[\boldsymbol{y}_t; \theta]$ is the mean of the process $\boldsymbol{y}$, and according to the spectral factorization theorem (see [120]), we can write $\boldsymbol{\zeta}_t = H(q,\theta)\boldsymbol{\varepsilon}_t$ in which $H(q,\theta)$ is a causal and causally and stably invertible LTI filter that can be parameterized (independently) by $\theta$, and $\boldsymbol{\varepsilon}$ is white noise, i.e.,

$$\boldsymbol{y}_t = \mathbb{E}[\boldsymbol{y}_t; \theta] + H(q,\theta)\boldsymbol{\varepsilon}_t, \quad t \in \mathbb{Z}. \tag{4.2}$$

It is clear that $\boldsymbol{\varepsilon}_t \in \boldsymbol{\mathcal{H}}_t$, and (4.2) is Wold's decomposition of $\boldsymbol{y}$.

Hence, this case is not much different from the case of LTI models in Example 2.4.2 (on page 35) where the mean of the process is $\mu_t(\theta) = G(q,\theta)u_t$. Regardless of how the (assumed known) mean is modeled, the process $\boldsymbol{y}$ has the same representation in terms of the mean signal and causally filtered (linear) innovations. Thus, we may proceed similarly and define two linear predictors. The first predictor can be constructed by ignoring the stochastic structure of the additive noise $\boldsymbol{\zeta}$ and carrying on as if $\boldsymbol{\zeta}$ were an independent white process. This leads to the deterministic predictor

$$\hat{y}_{t|t-1}(\theta) = \mu_t(\theta), \quad t \in \mathbb{Z}, \tag{4.3}$$

which is independent of the observed outputs. Under some conditions on the used inputs (e.g., open-loop operation) and $\mu_t$ as functions of $\theta$, the PEM estimator

$$\begin{aligned} \hat{\theta}(\boldsymbol{D}_N) &:= \arg\min_{\theta\in\Theta} \quad \sum_{t=1}^{N} \|\boldsymbol{e}_t(\theta)\|^2 \\ &\text{such that} \quad \boldsymbol{e}_t(\theta) = \boldsymbol{y}_t - \mu_t(\theta), \quad \forall t = 1, \ldots, N, \end{aligned}$$

can be shown to be consistent (see Example 2.4.4 for the linear case). However, it is obvious that the predictor in (4.2) is suboptimal, even in the restricted class of linear predictors. Note that it does not coincide with the unique optimal linear predictor which relies on the covariance function of $\boldsymbol{y}$ (see Theorem B.1.9

and Section B.4). The optimal "linear" (in $\boldsymbol{y}$) predictor can be constructed by inverting the noise model $H(q,\theta)$ (compare to (2.39)); we define

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) := [1 - H^{-1}(q,\theta)]\boldsymbol{y}_t + H^{-1}(q,\theta)\mu_t(\theta), \quad t \in \mathbb{Z}$$

where $\hat{\boldsymbol{y}}_{t|t-1}(\theta)$ denotes the predicted output at time $t$ given all outputs and inputs before $t$.

Now assume that the process $\boldsymbol{\zeta}$ is a purely non-deterministic non-stationary process with zero mean (compare to the linear state-space model in (2.45)). Then by Theorem 2.1.6, we can write

$$\boldsymbol{y}_t = \mu_t(\theta) + \sum_{k=0}^{\infty} h_k(t)\boldsymbol{\varepsilon}_{t-k}, \quad t \in \mathbb{Z},$$

in which $\boldsymbol{\varepsilon}$ is the (linear) innovations process of $\boldsymbol{y}$, and the linear filter is time-varying with a "square summable" impulse response sequence. Using the assumption of zero initial conditions; i.e., that both $\mu_t(\theta)$ and $\boldsymbol{\varepsilon}_t$ are zero for all $t < 1$, we have the vector representation

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{y}_3 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}}_{=\boldsymbol{Y}} = \underbrace{\begin{bmatrix} \mu_1(\theta) \\ \mu_2(\theta) \\ \mu_3(\theta) \\ \vdots \\ \mu_N(\theta) \end{bmatrix}}_{=:M(\theta)} + \underbrace{\begin{bmatrix} 1 & 0 & \dots & 0 \\ h_1(2) & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{N-1}(N) & h_{N-2}(N) & \dots & 1 \end{bmatrix}}_{=:H_N} \underbrace{\begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_N \end{bmatrix}}_{=:\boldsymbol{\mathcal{E}}}, \tag{4.4}$$

that is,

$$\boldsymbol{Y} = M(\theta) + H_N\boldsymbol{\mathcal{E}}. \tag{4.5}$$

Observe that $M(\theta)$ is a known deterministic vector-valued function of $\theta$, $H_N$ is a unitriangular matrix of deterministic (and decaying) coefficients, and the innovations vector $\boldsymbol{\mathcal{E}}$ has a zero mean and a finite covariance. It is then possible to define the best linear predictor of $\boldsymbol{Y}$ by inverting the matrix $H_N$.

Define the vector of stacked one-step ahead 'linear' predictors

$$\widehat{\boldsymbol{Y}} := \begin{bmatrix} \hat{\boldsymbol{y}}_{1|0}^{\top} & \dots & \hat{\boldsymbol{y}}_{N|N-1}^{\top} \end{bmatrix}^{\top}. \tag{4.6}$$

Then, similar to (2.40), we have

$$\widehat{\boldsymbol{Y}}(\theta) = [I - H_N^{-1}]\boldsymbol{Y} + H_N^{-1}M(\theta) \tag{4.7}$$

which is parameterized by $\theta$ and the coefficients of $H_N$. Observe that $H_N$ is invertible for any finite $N$; however, to be able to guarantee that $\widehat{\boldsymbol{Y}}$ is well defined as $N \to \infty$, it is required to assume that $\boldsymbol{y}$ is such that the entries of $H_N^{-1}$ are decaying as time grows (see Assumption 2.1.7 on page 20).

In the previous example, the models were given in the form of Wold's decomposition, and we showed how linear predictors (in $\boldsymbol{y}$) can be defined. In the next example, we will define five PEM instances for a relatively simple stochastic nonlinear model. The first step is to rewrite the given model in the form of Wold's decomposition; then the linear predictors and the corresponding PEM problems can be defined.

**Example 4.2.2** (PEM for a scalar stochastic nonlinear model)**.** Assume that the observations are generated according to the relation

$$\boldsymbol{y}_t = \theta(u_{t-1} + \boldsymbol{w}_t)^2 + \boldsymbol{\zeta}_t, \quad t = 1, \dots, N \tag{4.8}$$

in which all signals are scalars, $\theta = 0.7$, $u$ is a known input signal, $\boldsymbol{w}$ is an unobserved independent white noise with time-independent variance $\lambda_w = 1$, and $\boldsymbol{\zeta}$ is a linearly filtered white noise defined as

$$\boldsymbol{\zeta}_t := \frac{1}{1 - 0.9q^{-1}}\boldsymbol{\varepsilon}_t, \quad t \in \mathbb{Z}, \tag{4.9}$$

in which $\boldsymbol{\varepsilon}$ is a stationary white noise with variance $\lambda_\varepsilon = 3$. Observe that the full distributions of $\boldsymbol{w}$ and $\boldsymbol{\zeta}$ are not specified; however, it is assumed that $\boldsymbol{w}$ and $\boldsymbol{\zeta}$ are independent. Moreover, notice that the model is simplified in two ways: (i) the unobserved disturbance $\boldsymbol{w}$ is assumed to be an independent process, (ii) the model is linear in the parameter $\theta$. A straightforward extension of the model, that does not affect the discussion, is to assume a parameterized input (for example, $u_t(\theta) = G(q;\theta)\tilde{u}_t$ for some transfer operator $G$ and known $\tilde{u}_t$.)

Under the assumption that $\boldsymbol{\zeta}_0 = 0$, the model can be written in the vector form

$$\begin{aligned}\boldsymbol{Y} &= \theta(U + \boldsymbol{W})^2 + \boldsymbol{Z}\\ \boldsymbol{Z} &= H\boldsymbol{\mathcal{E}}\end{aligned}$$

in which the exponent is applied element-wise; i.e., for any vector $\boldsymbol{W}$ we define $\boldsymbol{W}^2 := [\boldsymbol{w}_1^2 \dots \boldsymbol{w}_N^2]^\top$. The vector $U$ is redefined, just for the current example, to be the vector $[u_0 \dots u_{N-1}]^\top$. According to (4.9), the filtering matrix

$$H = \begin{bmatrix} 1 & 0 & 0 & \dots & 0\\ 0.9 & 1 & 0 & \dots & 0\\ 0.9^2 & 0.9 & 1 & \dots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0.9^{N-1} & 0.9^{N-2} & \dots & \dots & 1 \end{bmatrix}.$$

The assumption that $\boldsymbol{w}$ is an independent white noise makes the model simple enough to allow for analytic computation of the outputs mean value. It holds that

$$\boldsymbol{Y} = \theta U^2 + \theta\boldsymbol{W}^2 + 2\theta U\boldsymbol{W} + \boldsymbol{Z},$$

$$\begin{aligned}\mu(U;\theta) &:= \mathbb{E}[\boldsymbol{Y};\theta] = \theta U^2 + \theta\mathbb{E}[\boldsymbol{W}^2] + 2\theta\mathbb{E}[U\boldsymbol{W}] + \mathbb{E}[\boldsymbol{Z}]\\ &= \theta(U^2 + \lambda_w\mathbf{1})\end{aligned}$$

in which the product $U\boldsymbol{W}$ is element-wise and $\mathbf{1} \in \mathbb{R}^N$ is a column vector of ones. Using this result, it is possible to evaluate the covariance matrix of the output vector, it holds that

$$\begin{aligned}\Sigma(U;\theta) &:= \mathbf{cov}(\boldsymbol{Y},\boldsymbol{Y};\theta)\\ &= \mathbb{E}\left[\boldsymbol{Y} - \mu(U;\theta))(\boldsymbol{Y} - \mu(U;\theta))^\top\right]\\ &= \mathbb{E}\left[(\theta\boldsymbol{W}^2 + 2\theta U\boldsymbol{W} + H\boldsymbol{\mathcal{E}} - \theta\lambda_w\mathbf{1})(\theta\boldsymbol{W}^2 + 2\theta U\boldsymbol{W} + H\boldsymbol{\mathcal{E}} - \theta\lambda_w\mathbf{1})^\top\right]\\ &= \mathbb{E}\left[\theta^2(\boldsymbol{W}^2)(\boldsymbol{W}^2)^\top + 4\theta^2(\boldsymbol{W}^2)(U\boldsymbol{W})^\top + 2\theta(\boldsymbol{W}^2)(H\boldsymbol{\mathcal{E}})^\top\right.\\ &\qquad - 2\theta^2\lambda_w(\boldsymbol{W}^2)(\mathbf{1})^\top + 4\theta^2(U\boldsymbol{W})(U\boldsymbol{W})^\top + 4\theta(U\boldsymbol{W})(H\boldsymbol{\mathcal{E}})^\top\\ &\qquad \left.-4\theta^2\lambda_w(U\boldsymbol{W})(\mathbf{1})^\top + H\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}^\top H^\top - 2\theta\lambda_w(H\boldsymbol{\mathcal{E}})(\mathbf{1})^\top + \theta^2\lambda_w^2 I\right]\end{aligned}$$

in which the products $(\cdot)(\cdot)^\top$ evaluate to diagonal matrices whose diagonal is given by the element-wise product of the vectors, and therefore

$$\Sigma(U;\theta) = \lambda_\varepsilon HH^\top + D(U;\theta) \tag{4.10}$$

in which $D$ is a diagonal matrix with entries $[D]_{tt} = 2\theta^2\lambda_w(2u_{t-1}^2 + \lambda_w)$ for all $t = 1,\dots N$. It is necessary to stress here that, due to the nonlinear dynamics, the matrix $D$ depends on the input (unlike the case of linear state-space models; see (2.25) and (2.26)). Furthermore, we define the matrix square-root

$$\Sigma^{\frac{1}{2}}(U;\theta) := L(U;\theta)\Lambda^{\frac{1}{2}}(U,\theta)$$

in which $L$ is the unique lower unitriangular diagonal matrix, and $\Lambda$ is a diagonal matrix.

Note that, for the current example, the mean vector is nonlinear in the known inputs. Also notice that the first term of the covariance matrix $\Sigma(U;\theta)$ as shown in (4.10) is due to the additive noise $\boldsymbol{Z}$ and is independent of $\theta$, while the second term is due to the "process noise" $\boldsymbol{W}$ and depends on the model (the nonlinearity), the variance $\lambda_w$ and the input signal.

The output can be written as

$$\boldsymbol{y}_t = \theta(u_{t-1}^2 + \lambda_w) + \sum_{k=0}^{t} l_k(t;U_t,\theta)\boldsymbol{\varepsilon}_{t-k}, \quad t \in \mathbb{N}, \tag{4.11}$$

in which $\boldsymbol{\varepsilon}_t$ is an innovation sequence; i.e., an uncorrelated sequence of zero mean random variables with time-dependent variances given by the diagonal of $\Lambda(U;\theta)$, and $l_t(0;U_t,\theta) := 1$ for all $t \in \mathbb{N}$. This representation of the outputs agrees with Wold's decomposition under the assumption of zero initial conditions. We note here that in a scenario where $\boldsymbol{\zeta}$ is an independent process, it holds that $H = I_N$, $L(U;\theta) = I_N$, and $\boldsymbol{y}$ is a linear process such that $\mathbb{E}[\boldsymbol{y}_t;\theta] = \mathbb{E}[\boldsymbol{y}_t|\mathcal{H}_{t-1};\theta]$.

We now define the following predictors:

$$\begin{aligned}
\widehat{Y}_1(\theta) &:= \theta U^2, \\
\widehat{Y}_2(\theta) &:= \theta(U^2 + \lambda_w), \\
\widehat{\boldsymbol{Y}}_3(\theta) &:= (I - \lambda_\varepsilon^{-\frac{1}{2}} H^{-1})\boldsymbol{Y} + \lambda_\varepsilon^{-\frac{1}{2}} H^{-1}\, \theta(U^2 + \lambda_w), \\
\widehat{\boldsymbol{Y}}_4(\theta) &:= (I - L^{-1}(U;\theta))\boldsymbol{Y} + L^{-1}(U;\theta)\, \theta(U^2 + \lambda_w),
\end{aligned} \tag{4.12}$$

and the following five PEM estimators

$$\begin{aligned}
\hat{\theta}_k(\boldsymbol{D}_N) &:= \arg\min_\theta \; \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_k(\theta)\|_2^2 \qquad \text{for } k = 1,2,3,4 \text{ and} \\
\hat{\theta}_5(\boldsymbol{D}_N) &:= \arg\min_\theta \; \left\{ \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_4(\theta)\|^2_{\Lambda^{-1}(U;\theta)} + \sum_{t=1}^N \log[\Lambda(U;\theta)]_{tt} \right\}.
\end{aligned} \tag{4.13}$$

The first two predictors are deterministic and both (as justification for their definition) assume that $\boldsymbol{y}$ is an independent process with unit variance; however, the first one ignores $\boldsymbol{w}$ completely. The third and fourth predictors are linear in $\boldsymbol{Y}$ and quadratic in $U$. Recall that both $H$ and $L^{-1}(U;\theta)$ are lower unitriangular matrices, and therefore the entries of the vectors $\widehat{\boldsymbol{Y}}_3$ and $\widehat{\boldsymbol{Y}}_4$ are one-step ahead predictors similar to (4.6).

To compare the five estimators defined in (4.13), we used simulated data according to the true model in (4.8). We generated 1000 independent realizations of $\boldsymbol{W}$ and $\boldsymbol{Z}$ for different values of $N$ between 100 and 3000, assuming Gaussian distributions with zero mean and the given variances. For each $N$ and each realization, the input $U$ is given as a known (fixed) realization of an independent Gaussian white noise with variance $\lambda_u = 5$.

The average MSE and the average bias of each estimator is given, for the different values of $N$, in Figure 4.1.

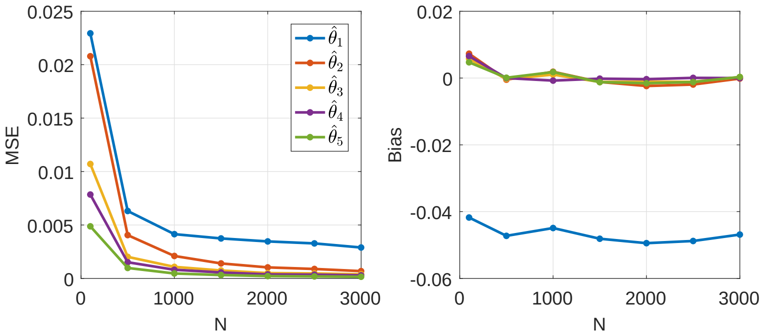


**Figure 4.1:** The average MSE and the average bias over 1000 Monte Carlo simulations of the five PEM estimators defined in (4.13).

The first observation to be made here is that the first estimator is biased. To understand why, observe that the predictor $\widehat{\boldsymbol{Y}}_1(\theta)$ ignores the process $\boldsymbol{w}$; i.e., it is defined assuming $\boldsymbol{w}_t = 0 \ \forall t$ and it is well-known that the PEM estimators based on this false assumption are biased (see [58]). Due to the convenient parameterization, the estimator has the closed-form expression

$$\hat{\theta}_1(\boldsymbol{D}_N) = \frac{\frac{1}{N}\sum_{t=1}^N \boldsymbol{y}_t u_{t-1}^2}{\frac{1}{N}\sum_{t=1}^N u_{t-1}^4}, \quad N \in \mathbb{N},$$

and it is possible to analyze the asymptotic estimate. By using the model in (4.8) to substitute for $\boldsymbol{y}_t$ in the above expression, we see that

$$\hat{\theta}_1(\boldsymbol{D}_N) = \theta^\circ \, \frac{\frac{1}{N}\sum_{t=1}^N u_{t-1}^4 + \frac{1}{N}\sum_{t=1}^N \boldsymbol{w}_t^2 u_{t-1}^2 + 2\frac{1}{N}\sum_{t=1}^N \boldsymbol{w}_t u_{t-1}^3}{\frac{1}{N}\sum_{t=1}^N u_{t-1}^4}.$$

Under some conditions on the input signal, a direct application of the law of large numbers together with Slutsky's lemma ([23, Chapter 5]) shows that

$$\hat{\theta}_1(\boldsymbol{D}_N) \xrightarrow{\text{a.s.}} \theta^\circ \left(1 + \lambda_w \frac{\lim\limits_{N\to\infty} \frac{1}{N}\sum_{t=1}^N u_{t-1}^2}{\lim\limits_{N\to\infty} \frac{1}{N}\sum_{t=1}^N u_{t-1}^4}\right) \quad \text{as} \quad N \to \infty.$$

and it is clear that $\hat{\theta}_1(\boldsymbol{D}_N) \xrightarrow{\text{a.s.}} \theta^* \neq \theta^\circ$ as $N \to \infty$ when $\lambda_w \neq 0$, and thus the estimator is not consistent. The asymptotic bias depends on the true parameter and the input properties. It is given by

$$\theta^\circ - \theta^* = -\theta^\circ \frac{\lambda_w}{3\lambda_u}$$

in which we used the fact that the input is Gaussian with variance $\lambda_u$. For the values used in the current example, this gives an asymptotic bias equal to $-0.7\frac{1}{15} = -0.04666$ which agrees with the simulation results in Figure 4.1.

On the other hand, we see that the simulation results indicate that all the other suggested estimators are consistent but have different accuracy. Observe for example that

$$\begin{aligned}\hat{\theta}_2(\boldsymbol{D}_N) &= \frac{\frac{1}{N}\sum_{t=1}^N \boldsymbol{y}_t(u_{t-1}^2 + \lambda_w)}{\frac{1}{N}\sum_{t=1}^N (u_{t-1}^2 + \lambda_w)^2} \\ &= \theta^\circ \frac{\frac{1}{N}\sum_{t=1}^N ((u_{t-1} + \boldsymbol{w}_t)^2 + \boldsymbol{\zeta}_t)(u_{t-1}^2 + \lambda_w)^2}{\frac{1}{N}\sum_{t=1}^N u_{t-1}^4 + \frac{1}{N}\sum_{t=1}^N \lambda_w^2 + \frac{2\lambda_w}{N}\sum_{t=1}^N u_{t-1}^2},\end{aligned}$$

and the asymptotic estimate exists; it holds that

$$\hat{\theta}_2(\boldsymbol{D}_N) \xrightarrow{\text{a.s.}} \theta^\circ \frac{\left[\lim\limits_{N\to\infty} \frac{1}{N}\sum_{t=1}^N u_{t-1}^4\right] + 2\lambda_w \left[\lim\limits_{N\to\infty} \frac{1}{N}\sum_{t=1}^N u_{t-1}^2\right] + \lambda_w^2}{\left[\lim\limits_{N\to\infty} \frac{1}{N}\sum_{t=1}^N u_{t-1}^4\right] + 2\lambda_w \left[\lim\limits_{N\to\infty} \frac{1}{N}\sum_{t=1}^N u_{t-1}^2\right] + \lambda_w^2} \quad \text{as} \quad N \to \infty,$$

and it is clear that $\hat{\theta}_2(\boldsymbol{D}_N) \xrightarrow{\text{a.s.}} \theta^\circ$ as $N \to \infty$ for every $\lambda_w > 0$ which shows that, unlike $\hat{\theta}_1$, $\hat{\theta}_2$ is consistent. Similar computations can be done to show the consistency of $\hat{\theta}_3$, $\hat{\theta}_4$ and $\hat{\theta}_5$.

It is of interest to observe that the four consistent PEM instances have prediction errors of the form

$$\boldsymbol{\mathcal{E}}_k = Q_k(\boldsymbol{Y} - \mu(U;\theta)) \quad k = 2,3,4, \text{ or } 5 \tag{4.14}$$

in which $Q_k$ is a square positive definite matrix (prefiltering/weighting);

$$Q_2 = I, \quad Q_3 = \lambda_\varepsilon^{-\frac{1}{2}} H^{-1}, \text{ and } \quad Q_4 = Q_5 = L^{-1}(U;\theta).$$

Regardless of the used $Q$, the three predictors deal with $\boldsymbol{w}$ in the same way: instead of ignoring it, they average over all possible values of $\boldsymbol{W}$ according to its known distribution. Furthermore, observe that since the weighted norm defining $\hat{\theta}_5$, see (4.13), is $\theta$-dependent, the objective function has a correction term that ensures the consistency of the estimator. We will discuss these observations further in the next section.

While Example 4.2.2 assumes a quite simple model, it illustrates the main ideas of this chapter. We saw that it is possible to define simple "distribution-independent predictors" that are linear in the observed outputs such that the resulting PEM estimator is consistent. The good news is that these ideas are still applicable for much more complicated models. In the next section, we give general definitions for linear predictors of nonlinear models and define several corresponding PEM estimators. We also clarify the definitions and behaviors of the estimators used in Example 4.2.2. However before moving there, we relate our framework to the notions of LTI second-order equivalent (LTI-SOE) models (see [36, 93]) and best linear approximations (BLA, see [111, 128, 130]).

### LTI second-order equivalent models and best linear approximations

Linear time-invariant approximations of nonlinear systems are usually considered under different set of assumptions and objectives. They are usually studied in a mean-square error framework where assumptions and restrictions on the systems to be approximated are implicitly given as assumptions on the input and output processes. It is usually assumed that the inputs and the outputs are zero mean stationary stochastic processes, such that the input belongs to a certain class; for example, a class of periodic processes, or processes that have a spectrum with a unique canonical spectral factorization.

In such a framework, explicit assumptions on the underlying data generating mechanism (such as a true parametric nonlinear model) are not necessarily used or required. The goal there is to use the assumptions on the data to obtain an LTI model – linear in both $\boldsymbol{y}$ and $\boldsymbol{u}$ – that approximate the behavior of the underlying nonlinear system. For example, an output error LTI-SOE model (OE-LTI-SOE) model is defined in [36, Section 4.2] as

$$G_{\mathrm{OE}}(q) := \arg \min_{G(q)\in\mathcal{G}} \mathbb{E}[\|\boldsymbol{y}_t - G(q)\boldsymbol{u}_t\|_2^2], \tag{4.15}$$

in which $\mathcal{G}$ is the set of all stable and causal LTI models and the expectation operator is with respect to the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{u}$. When the stability and causality constraints are dropped, the minimizer is called the best linear approximation (BLA) and denoted $G_{\mathrm{BLA}}(q)$. Observe that these models are not necessarily rational.

Apart from the obvious differences in the assumptions and the objectives, the obtained LTI-SOE models and BLAs depend on the distribution of the input process $\boldsymbol{u}$. Notice that the models in (4.15) are defined by averaging, not only over $\boldsymbol{y}$, but also over all inputs $\boldsymbol{u}$. Therefore, one has to speak of an LTI-SOE model or a BLA "with respect to a certain class of input signals". In this thesis, by contrast, the inputs are assumed fixed and known. They are used to describe the mean of $\boldsymbol{y}$, which is not necessarily stationary, and therefore all the computations are conditioned on the given inputs (consequently, the assumption regarding the underlying model structure is of importance here).

Even though linear predictors are utilized in this thesis to define PEM estimators, it should be clear that (under Assumption 2.2.1), the underlying models are exact once the full distribution of the basic stochastic process $\boldsymbol{\zeta}$ is known; i.e., the models completely specify the underlying measure $P_\theta$ in terms of all finite distributions of $\boldsymbol{y}$ for a given input. This comes in contrast to the undermodelling framework in which LTI-SOE models and BLAs are considered. For example, the OE-LTI-SOE (4.15) only captures the causal part of the cross-covariance function between $\boldsymbol{y}$ and $\boldsymbol{u}$, and the general error LTI-SOE (GE-LTI-SOE, see [36, Section 4.4]) only models the second-order properties of the signals in terms of the covariance function of $\boldsymbol{y}$ and the cross-covariance function between $\boldsymbol{y}$ and $\boldsymbol{u}$. To further clarify these remarks, we have the following example.

**Example 4.2.3** (LTI-SOEs model and BLAs)**.** Let us assume that the underlying nonlinear system is given by the relation

$$\boldsymbol{y}_t = \theta(\boldsymbol{u}_t + \boldsymbol{w}_t)^2 + \boldsymbol{v}_t - 2\theta, \quad t \in \mathbb{Z} \tag{4.16}$$

in which $\boldsymbol{u}$, $\boldsymbol{w}$ and $\boldsymbol{v}$ are independent and mutually independent zero mean stationary Gaussian processes with unit variance, $\theta \in \mathbb{R}$. It is straightforward to show that

$$\Phi_{\boldsymbol{u}}(z) = 1, \quad \Phi_{\boldsymbol{y}\boldsymbol{u}}(z) = 0, \quad \text{and} \quad \Phi_{\boldsymbol{y}}(z) = 8\theta^2 + 1.$$

Consequently, the OE-LTI-SOE model, with respect to the standard stationary Gaussian input, is $G_{OE}(q) = 0$, and the GE-LTI-SOE model is given by the triplet $(G_{\mathrm{GE}}(q), H_{\mathrm{GE}}(q), \lambda_0)$ in which $G_{\mathrm{GE}}(q) = 0$, $H_{\mathrm{GE}}(q) = 1$, and $\lambda_0 = 8\theta^2 + 1$ (see [36, Theorem 4.5]). The nonlinear system and its GE-LTI-SOE model are indistinguishable if only the second-order properties of $\boldsymbol{y}$ and $\boldsymbol{u}$ are considered, i.e., the process

$$\begin{aligned}\tilde{\boldsymbol{y}}_t &= G_{\mathrm{GE}}(q)\boldsymbol{u}_t + H_{\mathrm{GE}}(q)\tilde{\boldsymbol{\varepsilon}}_t,\\ &= H_{\mathrm{GE}}(q)\tilde{\boldsymbol{\varepsilon}}_t,\end{aligned} \tag{4.17}$$

in which $\tilde{\boldsymbol{\varepsilon}}_t$ is white noise with variance $\lambda_0$, has exactly the same spectrum as $\boldsymbol{y}$.

Observe that, because $\boldsymbol{y}$ is assumed to be a zero mean independent stationary process, the representation in (4.17) coincides with Wold's decomposition, see (2.5), in which $h_k = 0 \ \forall k \geq 1$. Moreover, under the assumptions on $\boldsymbol{u}$, it holds that the optimal linear predictor constructed based on (4.17) is equivalent to

$$\hat{y}_{t|t-1} = \mathbb{E}[\boldsymbol{y}_t] = \mathbb{E}[\boldsymbol{y}_t | \boldsymbol{\mathcal{H}_{t-1}}] = 0.$$

It can also be shown that the BLA (with respect to the standard stationary Gaussian input) as described in [131] or in [46], coincides with the OE-LTI-SOE model; i.e., $G_{\mathrm{BLA}}(q) = 0$.

On the other hand, the linear predictors suggested in this thesis are defined by conditioning on the assumed known (realization of the) input. Assuming that $u$ is fixed and known, the mean of $\boldsymbol{y}$ is

$$\mathbb{E}[\boldsymbol{y}_t; \theta] = \mathbb{E}[\theta(u_t + \boldsymbol{w}_t)^2 + \boldsymbol{v}_t - 2\theta] = \theta(u_t^2 - 1), \quad \forall t \in \mathbb{Z},$$

and the variance is

$$\mathbb{E}[(\boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t; \theta])^2; \theta] = 2\theta^2(2u_t^2 + 1) + 1 \quad \forall t \in \mathbb{Z},$$

and thus, Wold's decomposition of $\boldsymbol{y}$ (given $u$) is

$$\boldsymbol{y}_t = \theta(u_t^2 - 1) + H_t(q; \theta)\boldsymbol{\varepsilon}_t, \ \ \mathbf{var}(\boldsymbol{\varepsilon}_t) = 1, \quad \forall t \in \mathbb{Z}, \tag{4.18}$$

in which $H_t(q;\theta)$ is a time-varying filter whose impulse response coefficients $h_k(t) = 0 \ \forall k \geq 1$ and $h_0(t) = \sqrt{2\theta^2(2u_t^2 + 1) + 1}$ for all $t \in \mathbb{Z}$. Notice that, because $\boldsymbol{y}$ is an independent process, the optimal linear predictor (in $\boldsymbol{y}$) coincides with the unrestricted optimal MSE predictor as well as the unconditional mean:

$$\hat{y}_{t|t-1}(\theta) = \mathbb{E}[\boldsymbol{y}_t; \theta] = \mathbb{E}[\boldsymbol{y}_t | \boldsymbol{\mathcal{H}_{t-1}}; \theta] = \theta(u_t^2 - 1), \quad \forall t \in \mathbb{Z},$$

which is nonlinear in $u_t$.

Thus, a main difference between the models in (4.17) and (4.18) is the assumption on the input. While the GE-LTI-SOE model is defined by averaging over an assumed stationary input, the model in (4.18) is obtained by conditioning on a given realization.

## 4.3 Optimal Linear Predictors for Nonlinear Models

The general prediction problem can be described as follows: at time $t-1$, we have observed the outputs $\boldsymbol{y}_1, \dots, \boldsymbol{y}_{t-1}$ for some $t \in \mathbb{Z}$ and wish to estimate the value of $\boldsymbol{y}_t$. Let us define the column vector of outputs $\boldsymbol{Y}_{t-1} : [\boldsymbol{y}_1^\top, \dots, \boldsymbol{y}_{t-1}^\top]^\top \in \mathsf{L}_2^{d_y(t-1)}$. A one-step ahead predictor is defined as a (measurable) function of the observation vector $\boldsymbol{Y}_{t-1}$, which is usually chosen to minimize some criteria. We have already seen in Section 2.4.2 (on page 34) that the optimal predictor, in the sense of minimizing the Mean-Square Error (MSE), is the conditional mean given in (2.35) which is, in most cases, hard to compute. For this reason, and due to the fact that the prediction error framework does not really need an "optimal" predictor to construct a consistent estimator, we are led to the restricted class of linear predictors of the form

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) = f(U_t;\theta) + \sum_{k=1}^{t-1} \tilde{l}_{t-k}(t, U_t;\theta)\boldsymbol{y}_k, \quad t \in \mathbb{N}.$$

in which $f$ and $\tilde{l}_{t-k}$ are functions of $\theta$ and the known inputs.

Linear predictors are much easier to work with; notice for example that a unique optimal linear least MSE predictor for any second-order process always exists among the set of linear predictors (see below). The computations are also straightforward, and closed-form expressions for the predictors are available in several relevant cases.

### 4.3.1 The Optimal Linear Predictor

By considering the outputs of the general model in (2.10) as elements of the Hilbert space $\mathsf{L}_2^{d_y}$ (see Appendix B), the projection theorem can be used to define the optimal linear predictor. The key idea is that the optimal linear predictor can be thought of as the unique orthogonal projection of the random vector $\boldsymbol{y}_t$ onto the closed subspace spanned by the observation vector $\boldsymbol{Y}_{t-1}$ when the MSE is used as an optimality criteria. We formalize this in the following definition and lemma.

**Definition 4.3.1** (Optimal linear predictor)**.** *Let $\mathcal{S} \subset \mathsf{L}_2^{d_y}$ be the closed subspace spanned by the entries of $\boldsymbol{Y}_{t-1}$. Then, an optimal linear predictor of $\boldsymbol{y}_t$ in $\mathcal{S}$ is defined as a vector $\hat{\boldsymbol{y}}_{t|t-1} \in \mathcal{S}$ such that*

$$\|\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}\|^2_{\mathsf{L}_2^{d_y}} := \mathbb{E}\left[\|\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}\|_2^2;\theta\right] \leq \mathbb{E}\left[\|\boldsymbol{y}_t - \tilde{\boldsymbol{y}}\|_2^2;\theta\right] \quad \forall \tilde{\boldsymbol{y}} \in \mathcal{S}$$

A characterization of such an optimal predictor is given in the following lemma.

**Lemma 4.3.2.** *The optimal one-step ahead linear predictor defined in Definition 4.3.1 exists and is unique. It is given by*

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \mathbb{E}[\boldsymbol{y}_t;\theta] + \Psi(U_{t-1};\theta)\left(\boldsymbol{Y}_{t-1} - \mu(U_{t-1};\theta)\right), \quad 1 \leq t \leq N \tag{4.19}$$

*in which*

$$\mu(U_{t-1};\theta) := \mathbb{E}[\boldsymbol{Y}_{t-1};\theta],$$
$$\Psi(U_{t-1};\theta) := \left[\mathbf{cov}(\boldsymbol{Y}_{t-1}, \boldsymbol{Y}_{t-1};\theta)\right]^{-1}\mathbf{cov}(\boldsymbol{y}_t, \boldsymbol{Y}_{t-1};\theta), \tag{4.20}$$

*and $\hat{\boldsymbol{y}}_{1|0}(\theta) = \mathbb{E}[\boldsymbol{y}_1;\theta]$.*

*Proof.* The proof is given in Section B.4.2, by a direct application of Theorem B.1.9. ■

To connect this result to Wold's decomposition of $\boldsymbol{y}$ and its innovation process, observe that the predictors in (4.19) would be easy to compute if the matrices $\mathbf{cov}(\boldsymbol{Y}_{t-1}, \boldsymbol{Y}_{t-1}; \theta)$ were diagonal. This holds only if the outputs $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}$ are orthogonal (uncorrelated) which is rarely the case in most applications. Nevertheless, the Gram-Schmidt procedure (see [67, Section 4.2.3]) can be used to (causally) transform the observations into a set of orthogonal vectors (innovations) $\{\boldsymbol{\varepsilon}_k\}$ such that

$$\begin{aligned} \boldsymbol{\varepsilon}_t &:= \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta), \quad 1 \le t \le N, \\ &= \boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t; \theta] - \sum_{k=1}^{t-1} \mathbf{cov}(\boldsymbol{y}_t, \boldsymbol{\varepsilon}_k) \lambda_{\boldsymbol{\varepsilon}_k}^{-1} \boldsymbol{\varepsilon}_k, \end{aligned} \tag{4.21}$$

with $\boldsymbol{\varepsilon}_1 = \boldsymbol{y}_1 - \mathbb{E}[\boldsymbol{y}_1; \theta]$, and $\lambda_{\boldsymbol{\varepsilon}_k} = \mathbf{cov}(\boldsymbol{\varepsilon}_k, \boldsymbol{\varepsilon}_k)$.

Let

$$\boldsymbol{\mathcal{E}}_{t-1} = [\boldsymbol{\varepsilon}_1^\top \ldots \boldsymbol{\varepsilon}_{t-1}^\top]^\top, \quad \bar{\boldsymbol{Y}}_{t-1} := \boldsymbol{Y}_{t-1} - \mu(U_{t-1}; \theta) \quad \text{and,} \quad \bar{\boldsymbol{y}}_t := \boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t; \theta].$$

Then, for the purpose of linear prediction, the vectors $\boldsymbol{\mathcal{E}}_{t-1}$ and $\bar{\boldsymbol{Y}}_{t-1}$ are equivalent in the sense that they span the same subspaces and it holds that

$$\mathcal{P}_{\mathbf{sp}\{\bar{\mathbf{Y}}_{t-1}\}}[\bar{\boldsymbol{y}}_t] = \mathcal{P}_{\mathbf{sp}\{\boldsymbol{\mathcal{E}}_{t-1}\}}[\bar{\boldsymbol{y}}_t].$$

Consequently, under the assumption that all signals are known to be zero for all $t \le 0$, the above construction is identical to Wold's decomposition (see the second row of (2.6) and compare to (4.21)).

**Remark 4.3.3.** *Observe that the coefficients (4.20) in the expression of the optimal predictor depend only on the unconditional first and second moments of $\boldsymbol{y}$ up to time t. Consequently, the computations of the best linear predictor can be simpler than the computations of the unrestricted optimal predictor (the conditional mean).*

The next lemma concerns the computations of the optimal one-step ahead linear predictor. It shows that finding the optimal linear predictors and the associated innovations corresponds to an $LDL^\top$ factorization of the covariance of $\boldsymbol{Y}$ (see [51, Section 4.1] for the definition and the properties of such a factorization).

**Lemma 4.3.4** (Computations of the optimal linear predictor)**.** *Consider the general nonlinear model in (2.10) subject to Assumption 4.1.1 such that $\boldsymbol{y}_t = 0 \ \forall t \le 0$. Furthermore, assume that*

$$\mu(U; \theta) := \mathbb{E}[\boldsymbol{Y}; \theta], \quad \textit{and} \quad \Sigma(U; \theta) := \mathbf{cov}(\boldsymbol{Y}, \boldsymbol{Y}; \theta) \succ 0$$

*are given. Then the unique optimal one-step ahead linear predictor of $\boldsymbol{y}_t$ is given by*

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \mathbb{E}[\boldsymbol{y}_t; \theta] + \sum_{k=1}^{t-1} \tilde{l}_{t-k}(t, U_{t-1}; \theta) \left(\boldsymbol{y}_k - \mathbb{E}[\boldsymbol{y}_k; \theta]\right), \tag{4.22}$$

*in which*

$$\tilde{l}_j(t, U_{t-1}; \theta) := \left[L^{-1}(U;\theta)\right]_{tj}, \quad 1 < t \le N,$$

*and the matrix* $L(U;\theta)$ *is the unique lower unitriangular matrix given by the* $LDL^{\mathsf{T}}$ *factorization of* $\Sigma$*; that is,*

$$\Sigma(U;\theta) = L(U;\theta)\Lambda(U;\theta)L^{\mathsf{T}}(U;\theta). \tag{4.23}$$

*Furthermore,* $\hat{\boldsymbol{y}}_{1|0}(\theta) = \mathbb{E}[\boldsymbol{y}_1;\theta]$.

*Proof.* The last statement of the theorem is straightforward: given no observations (or zero initial condition), the subspace $\mathcal{S}$ is equal to $\{0\}$ because the span of an empty set is the zero vector. Therefore, the orthogonal projection of $\boldsymbol{y}_1 - \mathbb{E}[\boldsymbol{y}_1;\theta]$ onto $\mathcal{S}$ is 0 (see properties of projections in Appendix B) and the result follows.

To establish (4.22), first recall that whenever the covariance matrix $\Sigma$ is positive definite, the factorization in (4.23) is unique. Then observe that using Wold's decomposition or (4.21) we may write

$$\boldsymbol{Y} = \mu(U;\theta) + \tilde{L}(U;\theta)\boldsymbol{\mathcal{E}} \tag{4.24}$$

in which

$$\tilde{L}(U;\theta) = \begin{bmatrix} I & 0 & \dots & 0 \\ \mathbf{cov}(\boldsymbol{y}_2, \varepsilon_1)\lambda_{\varepsilon_1}^{-1} & I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{cov}(\boldsymbol{y}_N, \varepsilon_1)\lambda_{\varepsilon_1}^{-1} & \mathbf{cov}(\boldsymbol{y}_N, \varepsilon_2)\lambda_{\varepsilon_2}^{-1} & \dots & I \end{bmatrix}.$$

From (4.24) and due to the linearity of the expectation operator, it follows that

$$\mathbf{cov}(\boldsymbol{Y}, \boldsymbol{Y}; \theta) = \tilde{L}(U;\theta)\tilde{\Lambda}(U;\theta)\tilde{L}^{\mathsf{T}}(U;\theta).$$

The uniqueness of the factorization in (4.23) implies that $\tilde{L}(U;\theta) = L(U;\theta)$ and $\tilde{\Lambda}(U;\theta) = \Lambda(U;\theta)$ is a block diagonal matrix of innovation covariances.

Finally, observe that we are able to compute the innovations vector by inverting the unitriangular matrix $L$ (which is always invertible for any finite $N$) to get

$$\boldsymbol{\mathcal{E}}(\theta) = L^{-1}(U;\theta)(\boldsymbol{Y} - \mu(U;\theta)), \tag{4.25}$$

and by definition (see (4.21)) we have

$$\boldsymbol{\mathcal{E}}(\theta) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\theta), \tag{4.26}$$

which gives the vector of one-step ahead predictors

$$\widehat{\boldsymbol{Y}}(\theta) = \boldsymbol{Y} - \boldsymbol{\mathcal{E}}(\theta) = \boldsymbol{Y} - L^{-1}(U;\theta)(\boldsymbol{Y} - \mu(U;\theta)) \tag{4.27}$$

and therefore (4.22) holds. ∎

**Remark 4.3.5.** *Notice that (4.27) can be rewritten in the form*

$$\widehat{\boldsymbol{Y}}(\theta) = (I - L^{-1}(U;\theta))\boldsymbol{Y} + L^{-1}(U;\theta)\mu(U;\theta), \tag{4.28}$$

*which has the same form as the vector of optimal predictors of the linear case in (2.40) and (2.41). There, the entries of the matrix L are given by the impulse response coefficients of the noise model. Observe that here – unlike the linear case – the entries of $L^{-1}$ depend on the used input.*

**Example 4.3.1.** Consider the model of Example 4.2.2. The vector of optimal one-step ahead linear predictors is given by $\widehat{\boldsymbol{Y}}_4(\theta)$ in (4.12).

In the next part, we relate the above results to the two special cases of stationary and quasi-stationary signals.

### The stationary and the quasi-stationary cases

In the special case where $\boldsymbol{y}$ is a stationary (centered) process observed from $t = -\infty$ with autocovariance function $s_k = \mathbf{cov}(\boldsymbol{y}_t, \boldsymbol{y}_{t+k})$ and $t, k \in \mathbb{Z}$, the optimal one-step ahead linear predictor is given by the Wiener filter (or recursively by the Kalman filter) and the solution is equivalent to computing the canonical factorization of the spectrum $\Phi_y(z)$ of the process (see [2, Section 9.5] or [67, Chapter 7]).

Let $\Phi_y(z)$ be a strictly positive rational spectrum of a scalar stationary process $\boldsymbol{y}$. Then it holds that

$$\Phi(z) = L(z)\Lambda_\varepsilon L^T(z^{-1}) = \mathcal{Z}[s_k] \tag{4.29}$$

in which $\mathcal{Z}[s_k]$ denotes the $z$-transform of the auto-covariance sequence, and

$$L(z) = 1 + \sum_{k=1}^{\infty} l_k z^{-k} \tag{4.30}$$

is a minimum-phase causal LTI filter known as the spectral factor of $\boldsymbol{y}$. Hence, $L(z)$ has a well defined causal and stable LTI inverse: $L^{-1}(z) = 1 + \sum_{k=1}^{\infty} \tilde{l}_k z^{-k}$. The optimal one-step ahead predictor is then given by the pure prediction Wiener filter (see [67, Section 7.6.1]),

$$\hat{\boldsymbol{y}}_{t|t-1} = \left(1 - L^{-1}(q)\right)\boldsymbol{y}_t, \quad t \in \mathbb{Z}, \tag{4.31}$$

and the innovation process is given by

$$\boldsymbol{\varepsilon}_t = L^{-1}(q)\boldsymbol{y}_t, \quad t \in \mathbb{Z}. \tag{4.32}$$

Under the assumption of zero initial conditions, that is $\boldsymbol{y}_t = 0 \ \forall t \le 0$, only the first $t-1$ coefficients of the impulse response of $L^{-1}(q)$ are used to compute $\hat{\boldsymbol{y}}_{t|t-1}$. Due to the uniqueness of the optimal linear predictor, this implies that for a stationary $\boldsymbol{y}$ the matrix $L$ in Lemma 4.3.4 is a Toeplitz lower unitriangular matrix whose entries are given by the first $t-1$ impulse response coefficients of the spectral factor (4.30). We summarize this in the following corollary.

**Corollary 4.3.6.** *Assume that $\boldsymbol{y}$ is a second-order weakly stationary discrete-time stochastic process with a strictly positive rational spectrum and zero initial conditions. Then the matrix $L$ defining the optimal one-step ahead linear predictors in Lemma 4.3.4 has a Toeplitz structure and its first column coincides with the first $N$ coefficients of the spectral factor of $\boldsymbol{y}$.*

*Proof.* See above. ■

Similar arguments can be used for the quasi-stationary case. Consider for instance the quasi-stationary output of Example 2.4.2. Comparing the predictors as given in (2.39) and (2.40) to Lemma 4.3.4, we conclude that $L$ has a Toeplitz structure and its first column coincides with the first $N$ coefficients of the impulse response of the noise model $H(q)$.

For a general case of a non-stationary output, $L$ will not be Toeplitz; however, as can be seen from the discussion after Lemma 4.3.2, the entries of its rows correspond to the (square summable) impulse response of a time-varying filter. The $i^{th}$ row correspond to the first $i$ coefficients of the impulse response at time $t = i$. The question of how to compute these time-varying filters in terms of a finite parameterization is not trivial though. Moreover, as we show in Section 4.4, assumptions have to be imposed on the predictors to establish the convergence of the PEM.

### 4.3.2 PEM Based on the Optimal Linear Predictor

We now define two instances of the PEM based on the unique optimal linear predictor. The difference between the two is due to the used criterion $\ell$ that maps the sequence of prediction errors into a positive scalar. The analysis of the resulting estimators is given in Section 4.4. The first instance is based on an unweighted time- and parameter-independent Euclidean norm.

**Definition 4.3.7** (PEMs based on the optimal linear predictor (L-PEM))**.** *A PEM estimator based on the optimal linear predictor is defined by*

$$\begin{aligned} \hat{\theta}(\boldsymbol{D}_N) = \arg\min_{\theta\in\Theta} \quad & \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\theta)\|_2^2 \\ \textit{such that} \quad & \widehat{\boldsymbol{Y}}(\theta) = (I - L^{-1}(U;\theta))\boldsymbol{Y} - L^{-1}(U;\theta)\mu(U;\theta). \end{aligned} \tag{4.33}$$

We will refer to this estimator as L-PEM (Linear Prediction Error Method) to indicate that the optimal linear predictor is used with a quadratic criterion. Observe that such an estimator was used in Example 4.2.2 in the previous section to define the consistent estimator $\hat{\theta}_4(\boldsymbol{D}_N)$. Also observe that in the classical case of LTI models, the L-PEM estimator is nothing more than the commonly used PEM estimator defined using a Euclidean norm and the optimal one-step ahead predictor (see Example 2.4.2).

To gain insight into the average behavior of the L-PEM estimators, observe that the objective function can be written in the form

$$\mathbf{tr}\left(L^{-\top}(U;\theta)L^{\top}(U;\theta)(\boldsymbol{Y}\boldsymbol{Y}^{\top} - 2\mu(U;\theta)\boldsymbol{Y}^{\top} + \mu(U;\theta))\right).$$

Integrating with respect to the true distribution of $\boldsymbol{Y}$ (i.e., the true measure $P_{\theta^\circ}$) and differentiating with respect to $\theta$ we get

$$\begin{aligned}&-2\mathbf{tr}\left(L^{-1}(U;\theta)\partial_\theta L(U;\theta)\Lambda(U;\theta^\circ)\right)\\&\quad-2\mathbf{tr}\left(L^{-\top}(U;\theta)L^{-1}(U;\theta)\partial_\theta L^{-1}(U;\theta)\Lambda(U;\theta^\circ)\right)L^{-1}(U;\theta)M)\\&\qquad-2\mathbf{tr}\left(L^{-\top}(U;\theta)L^{-1}(U;\theta)(\partial_\theta\mu(U;\theta)\mu(U;\theta^\circ)-\partial_\theta\mu(U;\theta)\mu(U;\theta)\right)\end{aligned}\tag{4.34}$$

in which

$$M=\mu(U;\theta^\circ)\mu^\top(U;\theta^\circ)-2\mu(U;\theta)\mu^\top(U;\theta^\circ)+\mu(U;\theta)\mu^\top(U;\theta).$$

It is easy to verify that this quantity vanishes if and only if $\mu(U;\theta)=\mu(U;\theta^\circ)$ and $L(U;\theta)=L(U;\theta^\circ)$. Note for example that the first term of (4.34) is zero, since $L^{-1}$ is lower unitriangular, $\partial_\theta L$ is strictly lower triangular (lower triangular with 0s along the main diagonal) and $\Lambda$ is diagonal. Thus, under identifiability and regularity conditions, the L-PEM estimator can be shown to be consistent. Note that in the linear case, the above condition is nothing more than the identifiability condition on the plant model $G(q,\theta)$ and the noise model $H(q,\theta)$.

The asymptotic properties of the L-PEM estimators can be improved by the use of a weighted criterion. To motivate this, observe that in the case of non-stationary linear Gaussian state-space models (see (2.48) and (2.49) in Example 2.4.3), the (unweighted) L-PEM is suboptimal, and using a weighted criterion according to the Gaussian distribution of the innovations results in an asymptotically efficient (ML) estimator. This leads us to the following PEM estimator.

**Definition 4.3.8** (PEM based on the optimal linear predictor using a weighted norm (WL-PEM))**.** *The weighted PEM estimator based on the optimal linear predictor and $\theta$-dependent weights is defined by*

$$\begin{aligned}\hat\theta(\boldsymbol{D}_N)=\arg\min_{\theta\in\Theta}\quad&\|\boldsymbol{Y}-\widehat{\boldsymbol{Y}}(\theta)\|^2_{\Lambda^{-1}(U;\theta)}+\log\det\Lambda(U;\theta)\\\textit{such that}\quad&\widehat{\boldsymbol{Y}}(\theta)=(I-L^{-1}(U;\theta))\boldsymbol{Y}-L^{-1}(U;\theta)\mu(U;\theta).\end{aligned}\tag{4.35}$$

We will refer to this estimator as WL-PEM (Weighted Linear Prediction Error Method) to indicate that the optimal linear predictor is used with a weighted quadratic criterion plus a log det term. Observe that the used criterion is on the "Gaussian log-likelihood function" form; it is both time- and parameter-dependent via the innovation covariance matrices. Also notice that the resulting estimator was used in Example 4.2.2 to define the consistent estimator $\hat\theta_5(\boldsymbol{D}_N)$. As indicated by the simulation results, the accuracy of the WL-PEM estimator is the best among the tested estimators in that example.

Note that, by using (4.25) and (4.26), the objective function in (4.35) can be equivalently written as

$$(\boldsymbol{Y}-\mu(U;\theta))^\top L^{-\top}(U;\theta)\Lambda^{-1}(U;\theta)L^{-1}(U;\theta)(\boldsymbol{Y}-\mu(U;\theta))+\log\det\Lambda(U;\theta)$$

Recalling (4.23) and that $L(U;\theta)$ is a unitriangular matrix, this expression is equivalent to

$$(\boldsymbol{Y} - \mu(U;\theta))^{\top}\Sigma^{-1}(U;\theta)(\boldsymbol{Y} - \mu(U;\theta)) + \log\det\Sigma(U;\theta)$$

in which $\Sigma(U,\theta)$ is the covariance of $\boldsymbol{Y}$. Observe that the $\log\det$ term of the criterion function in (4.35) is important for the consistency of the estimator due to the dependence of the weight on $\theta$. To see this, observe that for all $N$, every solution $\hat{\theta}$ to the problem in (4.35) must satisfy the equation

$$\left\{\partial_\theta\left[(\boldsymbol{Y} - \mu(U;\theta))^{\top}\Sigma^{-1}(U,\theta)(\boldsymbol{Y} - \mu(U;\theta)) + \log\det\Sigma(U;\theta)\right]\right\}\Big|_{\theta=\hat{\theta}} = 0$$

which can be expanded to

$$\begin{aligned}\big\{&\mathbf{tr}(\Sigma^{-1}(U;\theta)\partial_\theta\Sigma(U;\theta)\Sigma^{-1}(U;\theta)(\boldsymbol{Y} - \mu(U;\theta))(\boldsymbol{Y} - \mu(U;\theta))^{\top})\\ &+ \mathbf{tr}(\Sigma^{-1}(U;\theta)\partial_\theta\Sigma(U;\theta)) - 2\partial_\theta\mu^{\top}(U;\theta)\Sigma^{-1}(U;\theta)\boldsymbol{Y}\\ &+\ 2\partial_\theta\mu^{\top}(U;\theta)\Sigma^{-1}(U;\theta)\mu^{\top}(U;\theta)\big\}\big|_{\theta=\hat{\theta}} = 0.\end{aligned}$$

Taking the expectation with respect to the true distribution of $\boldsymbol{Y}$, we get that

$$\begin{aligned}&\big\{\mathbf{tr}(\Sigma^{-1}(U;\theta)\partial_\theta\Sigma(U;\theta)\Sigma^{-1}(U;\theta)\Sigma(U;\theta^{\circ})) + \mathbf{tr}(\Sigma^{-1}(U;\theta)\nabla_\theta\Sigma(U;\theta))\\ &-2\partial_\theta\mu^{\top}(U;\theta)\Sigma^{-1}(U;\theta)\mu(U;\theta^{\circ}) + 2\partial_\theta\mu^{\top}(U;\theta)\Sigma^{-1}(U;\theta)\mu(U;\theta)\big\}\big|_{\theta=\hat{\theta}} = 0\end{aligned}$$

which holds if and only if $\mu(U;\hat{\theta}) = \mu(U;\theta^{\circ})$ and $\Sigma(U;\hat{\theta}) = \Sigma(U;\theta^{\circ})$. If the model is identifiable, this can happen only when $\hat{\theta} = \theta^{\circ}$. Removing the $\log\det$ term will therefore result in a biased estimator.

In cases where the covariance is assumed to be known and independent of $\theta$, the $\log\det$ term will not be needed. See for example the definition of $\widehat{\boldsymbol{Y}}_3(\theta)$ and $\hat{\theta}_3(\boldsymbol{D}_N)$ of Example 4.2.2. An alternative way that can be used to avoid the $\log\det$ term when the required weights are parameter-dependent is to use a two-step procedure. First, an unweighted problem defining a consistent estimator is solved. Then, the resulting estimate is used to compute a consistent estimator of the weights, which is then used in the weighted problem.

Next, we consider the problem of constructing a suboptimal linear predictor. Here, we are thinking of the simplest possible predictor that can be used to still obtain consistent estimators of $\theta$.

### 4.3.3 Suboptimal Linear Predictors

Due to the uniqueness of the optimal linear predictor, any other linear predictor is suboptimal. Motivated by the Output-Error predictor in (2.42) which was shown to be consistent in the case of LTI models operating in open-loop (see Example 2.4.4), we define the following suboptimal predictor for general nonlinear models.

**Definition 4.3.9** (OE-type linear predictors). *Consider the general model given in (2.10). The Output-Error-type (OE-type) one-step ahead linear predictor of $\boldsymbol{y}_t$ is defined as the "deterministic" quantity*

$$\hat{y}_{t|t-1}(\theta) := \mathbb{E}[\boldsymbol{y}_t;\theta], \quad t \in \mathbb{N}. \tag{4.36}$$

*When Assumption 4.1.1 holds and $\boldsymbol{V}$ has a zero mean, the above predictors are given in the vector form*

$$\widehat{Y}(\theta) := \mu(U;\theta) = \mathbb{E}\left[\mathcal{M}(U,\boldsymbol{W};\theta);\theta\right]$$

*in which the expectation is with respect to the PDF $p(\boldsymbol{W};\theta)$.*

Notice that we did not assume the knowledge of the full distribution of $\boldsymbol{v}$ but only that it has a zero mean. The PDF of $\boldsymbol{W}$ is assumed to be known but can be parameterized by $\theta$. In some specific cases, the computations of the suboptimal predictor require the knowledge of only the first $r$ moments of $\boldsymbol{W}$, for some finite $r \in \mathbb{N}$ (for example, the case of Wiener model with polynomial nonlinearity of order $r$). The predictor in (4.36), although independent of $\boldsymbol{Y}$, is different from the "nonlinear simulation model" defined in [92, Section 5.3, page 147]. The above OE-type predictor averages the model outputs over all possible values of the unobserved process (instead of ignoring it). Under some identifiability and regularity conditions, this is expected to give rise to a consistent PEM estimator, which we define next.

**Definition 4.3.10** (PEM based on OE-type linear predictor). *The PEM estimator based on the OE-type linear predictor (OE-PEM) is defined by*

$$\begin{aligned} \hat{\theta}(\boldsymbol{D}_N) = \arg\min_{\theta\in\Theta} \quad & \|\boldsymbol{Y} - \widehat{Y}(\theta)\|^2 \\ \textit{such that} \quad & \widehat{Y}(\theta) = \mu(U;\theta). \end{aligned} \tag{4.37}$$

Any solution $\hat{\theta}$ to the problem in (4.37) must satisfy the condition

$$\partial_\theta \|\boldsymbol{Y} - \mu(U;\theta)\|^2\big|_{\theta=\hat{\theta}} = 0.$$

Expanding the derivative, we see that

$$\begin{aligned} \partial_\theta \|\boldsymbol{Y} - \mu(U;\theta)\|^2 &= \partial_\theta \left(\boldsymbol{Y}^\top \boldsymbol{Y} - 2\boldsymbol{Y}^\top \mu(U;\theta) + \mu(U;\theta)\mu^\top(U;\theta)\right) \\ &= -2\boldsymbol{Y}^\top \partial_\theta \mu(U;\theta) + 2\mu^\top(U;\theta)\partial_\theta \mu(U;\theta). \end{aligned}$$

Applying the expectation operator, with respect to the true distribution of $\boldsymbol{Y}$, to both sides yields

$$\begin{aligned} \mathbb{E}\left[\partial_\theta \|\boldsymbol{Y} - \mu(U;\theta)\|^2;\theta^\circ\right] &= \mathbb{E}\left[-2\boldsymbol{Y}^\top \partial_\theta \mu(U;\theta) + 2\mu(U;\theta)\partial_\theta \mu(U;\theta);\theta^\circ\right] \\ &= -2\mu^\top(U;\theta^\circ)\partial_\theta \mu(U;\theta) + 2\mu(U;\theta)\partial_\theta \mu(U;\theta). \end{aligned}$$

Therefore, the condition

$$\mathbb{E}\left[\partial_\theta \|\boldsymbol{Y} - \mu(U;\theta)\|^2;\theta^\circ\right]\big|_{\theta=\hat{\theta}} = 0$$

holds if and only if

$$\mu(U;\theta^\circ) = \mu(U;\hat{\theta}).$$

Under identifiability conditions, this can only happen when $\hat{\theta} = \theta^\circ$. This argument indicates that, under the required identifiability condition, the estimator in Definition 4.3.10 can be shown to be a consistent estimator. Observe that this estimator was used in Example 4.2.2 in the previous section to define the consistent estimator $\hat{\theta}_2(\boldsymbol{D}_N)$. We also note here that the norm in (4.37) can be weighted using any positive definite matrix independent of $\theta$ (see (4.14)).

In the following example, we compare the OE-PEM and the WL-PEM estimators defined in (4.37) and (4.35) to the (asymptotically efficient) MLE. We will assume a SISO model in two scenarios. In the first, we will consider a model with independent outputs over time such that the likelihood function factorizes as shown in (3.51). This allows for the use of deterministic numerical integration for the MLE computations. Recall that in this scenario, as we remarked before, the optimal linear predictor coincides with the unrestricted optimal predictor (conditional mean); both are equal to the unconditional mean value. Therefore, here, the OE-PEM and the WL-PEM estimators are using the same predictors but different criterion. In the second scenario, we will consider a case with dependent outputs. Here, we will compare the OE-PEM to a PEM estimator ignoring $\boldsymbol{w}$.

**Example 4.3.2** (PEM for stochastic Wiener model)**.** Let the observations be the outputs of a first-order SISO stochastic Wiener model given by the relations

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{x}_t^2 + \boldsymbol{v}_t, & t &= 1,\ldots,N, \\ \boldsymbol{x}_t &= \frac{q^{-1}}{1-\theta q^{-1}} u_t + \boldsymbol{w}_t, & \theta &= 0.7, \end{aligned} \tag{4.38}$$

in which we assume that $u$ is a known signal, $\boldsymbol{w}$ is an independent Gaussian process with zero mean and variance $\lambda_w = 1$, and $\boldsymbol{v}$ is an independent Gaussian process with zero mean and variance $\lambda_v = 3$ and independent of $\boldsymbol{w}$. We will assume that the input $u$ is a known realization of an independent Gaussian process with zero mean and variance $\lambda_u = 3$. This model is very similar to the model of Example 4.2.2; however, here we assume that we know the distributions of $\boldsymbol{w}$ and $\boldsymbol{v}$ to be able to compute the MLE. Since the outputs are independent, it is possible to use deterministic numerical integration, the Gauss-Hermite quadrature for example, to approximate the likelihood function.

Figure 4.2 shows the result of a Monte Carlo simulation over 1000 independent realizations of the inputs, $\boldsymbol{W}$ and $\boldsymbol{V}$ for different values of $N$ between 100 and 3000. The quasi-Newton algorithm, initialized at the true parameter, was used to compute the estimates. As expected, the two PEM instances are consistent. Furthermore, for this example, their MSE follows closely the MSE of the MLE but the MSE of the WL-PEM estimator is slightly better compared to the OE-PEM estimator. Figure 4.3 compares the cost functions of the three estimators

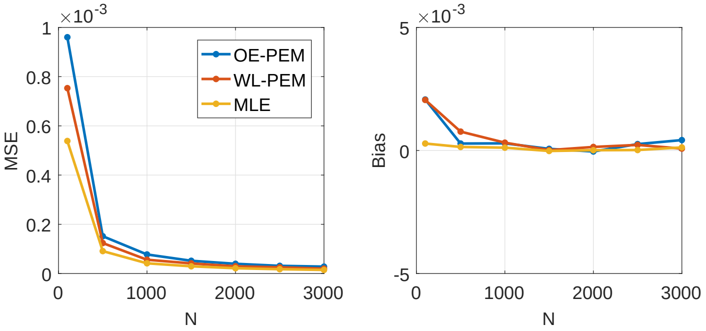


**Figure 4.2:** The average MSE and the average bias over 1000 Monte Carlo simulations of the OE-PEM and WL-PEM estimators defined in (4.37) (blue) and (4.35) (red) and the MLE (yellow) for the model in (4.38).

for a single realization when $N = 1000$. Notice that, for the model in (4.38), the cost functions of the PEM estimators are available in closed form, however the cost function of the MLE is not and it was approximated using the Gauss-Hermite quadrature. Consequently, the computations of the two PEM estimators are several times faster than the MLE, especially for large $N$.

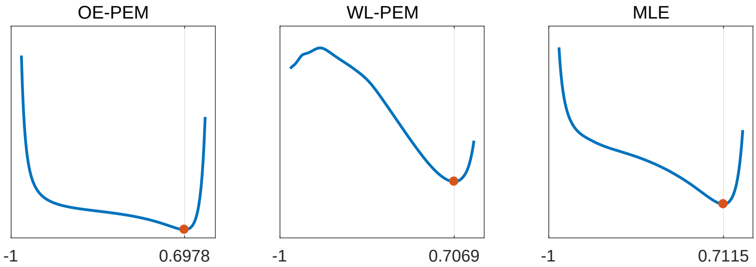


**Figure 4.3:** Sample of the cost functions of the three estimators in Example 4.3.2 for $N = 1000$. From left to right, the figure shows the cost function of (4.37), (4.35), and the approximated negative log-likelihood function.

Let us now assume that $\boldsymbol{w}$ is in fact colored; for example, consider

$$\boldsymbol{w}_t = \frac{q^{-1}}{1 - 0.9q^{-1}} \tilde{\boldsymbol{w}}_t, \quad t \in \mathbb{Z},$$

in which $\tilde{\boldsymbol{w}}$ is an independent Gaussian process with zero mean and unit variance. In this case, the output process $\boldsymbol{y}$ is colored and the computations of the MLE are challenging. While any predictor ignoring the process noise ([92, Section 5.3]) is known to lead to a biased PEM estimator, the OE-PEM estimator is a simple consistent estimator. By only assuming the stationarity of $\boldsymbol{w}$, we may write

$$\mathbb{E}[\boldsymbol{y}_t;\theta,\lambda_w] = \mu_t(\theta) + \lambda_w,$$
$$\mu_t(\theta) := \left(\frac{q^{-1}}{1-\theta q^{-1}}u_t\right)^2, \quad t \in \mathbb{Z}$$

and the OE-PEM estimator is defined by

$$\hat{\theta} := \arg\min_{\theta,\lambda_w} \sum_{t=1}^{N}(y_t - \mathbb{E}[\boldsymbol{y}_t;\theta,\lambda_w])^2, \tag{4.39}$$

in which we also optimize over $\lambda_w$. Figure 4.4 shows the result of a Monte Carlo simulation over 1000 independent realizations of the inputs, the colored $\boldsymbol{W}$ and $\boldsymbol{V}$ for different values of $N$ between 100 and 5000.

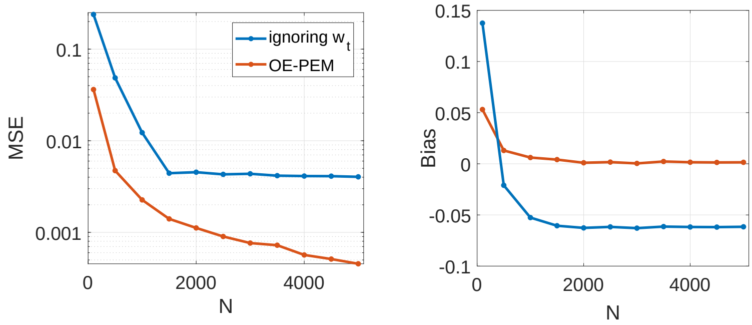


**Figure 4.4:** The average MSE (log-lin scale) and the average bias over 1000 Monte Carlo simulations when the disturbance is colored; the OE-PEM estimator defined in (4.39) is shown in blue, and an estimator based on a simulation predictor ignoring $\boldsymbol{w}$ is shown in red.

---

The results of the above example are encouraging. Recall, from Chapter 3, that approaches targeting the MLE for similar models were either without guarantees or computationally troublesome, see for instance Example 3.3.3 on page 61. On the contrary, using the PEM ideas suggested in this chapter leads to *closed-form cost functions* for several models with intractable likelihood functions.

For example, this will be the case for the class of stochastic Wiener models whenever the static nonlinearity is a polynomial; that is, when the model is given by the relation

$$\begin{aligned}\boldsymbol{y}_t &= f(\boldsymbol{x}_t;\theta) + \boldsymbol{v}_t, \quad t = 1,\ldots,N\\ \boldsymbol{x}_t &= G(q,\theta)u_t + \boldsymbol{w}_t,\end{aligned} \tag{4.40}$$

where $f$ is a multivariate polynomial. In this case, the mean of the model outputs is an explicit function of the moments of $\boldsymbol{w}$, which can be either evaluated or used as decision variables in the optimization problem. This is a great advantage that simplifies the computations and is expected to reduce the computational time significantly.

The next example clarifies this point using a more complicated model.

**Example 4.3.3** (Stochastic Wiener-Hammerstein Models)**.** Consider the model, depicted in Figure 4.5 given by the relations

$$\begin{aligned}\boldsymbol{y}_t &= G_2(q,\theta)f(y_t^{(1)}(\theta)+\boldsymbol{w}_t)+\boldsymbol{v}_t,\\ y_t^{(1)}(\theta) &= G_1(q,\theta)u_t,\\ \boldsymbol{w}_t &= H_1(q)\boldsymbol{\varepsilon}_t^{(2)},\\ \boldsymbol{v}_t &= H_2(q)\boldsymbol{\varepsilon}_t^{(2)} \quad t=1,\dots,N\end{aligned}$$

in which $\boldsymbol{\varepsilon}_t^{(1)}$ and $\boldsymbol{\varepsilon}_t^{(2)}$ are stationary white noises. The function $f(\cdot)$ denotes a static nonlinearity, and $G_1(q,\theta)$, $G_2(q,\theta)$ are rational transfer operators. The processes $\boldsymbol{w}$ and $\boldsymbol{v}$ are unobserved colored stationary process disturbance and measurement noise respectively. Therefore, methods relying on the likelihood function are analytically intractable. However, with minimal assumptions, a consistent estimator can be constructed by considering the PEM in (4.37).

To look at a concrete case, assume that all the signals are scalars, the static nonlinearity is $f(x) := x^2\ \forall x \in \mathbb{R}$, and recall that $\boldsymbol{w}$ is stationary. Then define

$$\begin{aligned}\hat{y}_{t|t-1}(\theta) &= \mathbb{E}\left[G_2(q,\theta)f(y_t^{(1)}(\theta)+\boldsymbol{w}_t)+H_2(q)\boldsymbol{\varepsilon}_t^{(2)}\right]\\ &= G_2(q,\theta)\mathbb{E}\left[f(y_t^{(1)}(\theta)+\boldsymbol{w}_t)\right]\\ &= G_2(q,\theta)\mathbb{E}\left[[y_t^{(1)}(\theta)]^2+\boldsymbol{w}_t^2+2y_t^{(1)}(\theta)\boldsymbol{w}_t)\right]\\ &= G_2(q,\theta)([y_t^{(1)}(\theta)]^2+\lambda_w).\end{aligned} \tag{4.41}$$

The consistent OE-PEM estimator is given by the minimization of a closed-form cost function; using (4.41) it holds that

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta,\lambda_w} \sum_{t=1}^{N}\left(\boldsymbol{y}_t - G_2(q,\theta)([G_1(q,\theta)u_t]^2+\lambda_w)\right)^2, \tag{4.42}$$

Notice that the unobserved process $\boldsymbol{v}$ is related linearly to the outputs; consequently the estimator in (4.42) can be easily improved by assuming a noise model for $\boldsymbol{v}$. Let us assume that $\boldsymbol{v}$ is stationary with a rational and strictly positive spectrum (such that $H_2$ can be modeled by a rational function) and append the new parameters to $\theta$. It is then possible to define the PEM estimator

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta,\lambda_w} \sum_{t=1}^{N} H_2^{-1}(q;\theta)\left(\boldsymbol{y}_t - G_2(q,\theta)([G_1(q,\theta)u_t]^2+\lambda_w)\right)^2, \tag{4.43}$$

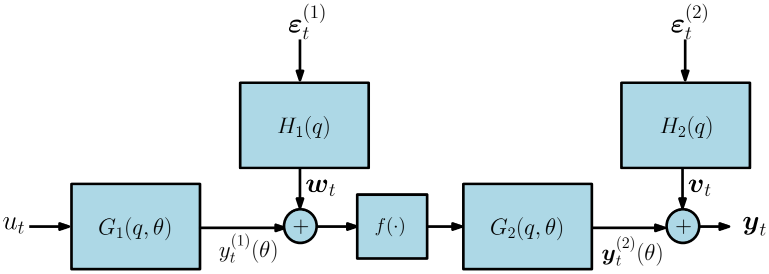


**Figure 4.5:** A stochastic Wiener-Hammerstein model. The LTI blocks are modeled with rational transfer operators $G_1(q,\theta)$ and $G_2(q,\theta)$. The function $f(\cdot)$ denotes a static nonlinearity. The processes $\boldsymbol{w}$ and $\boldsymbol{v}$ are colored unobserved stationary process disturbance and measurement noise respectively.

in which $H_2(q,\theta)$ might be known or independently parameterized by $\theta$. We note here that this estimator is similar to the third estimator estimator $\hat{\theta}_3$, see (4.12) and (4.13), defined in Example 4.2.2. Observe that the only assumption we used regarding $\boldsymbol{w}$ and $\boldsymbol{v}$ was stationarity and nothing else.

To look at a simulation example, consider the case where

$$G_1(q,\theta) = \frac{q^{-1}}{1-\theta_1 q^{-1}}, \qquad G_2(q,\theta) = \frac{q^{-1}}{1-\theta_2 q^{-1}},$$
$$H_1(q) = \frac{q^{-1}}{1-0.9q^{-1}}, \qquad H_2(q) = \frac{q^{-1}}{1-0.7q^{-1}},$$

and let $\theta_1 = 0.7$, $\theta_2 = 0.5$, $\varepsilon_t^{(1)} \sim \mathcal{N}(0,1)$ independent over $t$, and $\varepsilon_t^{(2)} \sim \mathcal{N}(0,3)$ independent over $t$ and independent of $\boldsymbol{\varepsilon}^{(1)}$ for all $t$. Let the input be a realization of $\boldsymbol{u}_t \sim \mathcal{N}(0,3)$ independent over time and independent of the disturbance and the noise. Assume that the noise model $H_2(q)$ is known, and consider the estimation problem of $\theta_1$ and $\theta_2$.

Figure 4.6 shows the result of a Monte Carlo simulation over 1000 independent realizations of the inputs, $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ for different values of $N$ between 100 and 3000. As expected, it is clear that both estimators are consistent. Moreover, it is obvious that using the noise model improves the accuracy of the estimator.

## 4.4 Asymptotic Analysis

Ideally, one would like to obtain exact information regarding the accuracy of the estimator when a data record of finite length is used. However, except in very few situations (such as the case of linear least-squares estimators), the analysis of frequentist estimators based on finite data records is intractable. On the other hand, it is usually possible to obtain precise asymptotic properties of the estimation method as $N \to \infty$. This is the objective of this section.

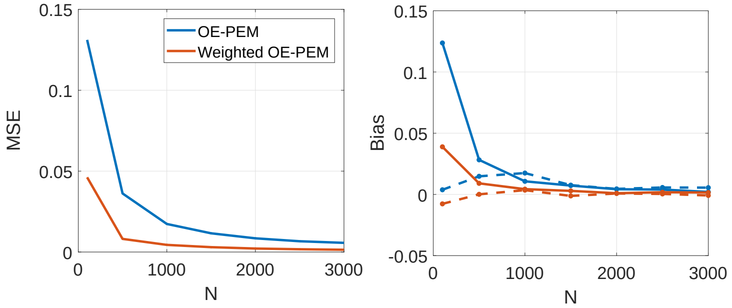


**Figure 4.6:** The average MSE and the average bias over 1000 Monte Carlo simulations when the disturbance is colored;The OE-PEM estimator defined in (4.42) is shown in blue, and the weighted version defined in (4.43) is shown in red. On the right, the solid lines correspond to $\theta_1$, and the dashed lines correspond to $\theta_2$

Besides providing a theoretical justification and confidence in the used estimation methods, the results of asymptotic[3] analysis are usually used to compare estimators (in terms of asymptotic variances or rates of convergence for example), and evaluate the accuracy of the estimates. The main disadvantage of this type of analysis is the lack of any guarantees when $N$ is finite. For example, consider an estimator that returns a constant value regardless of the data as long as $N < 10^6$ but is equal to the ML estimator for any larger $N$. Asymptotic analysis does not distinguish between the two; they are asymptotically equivalent.

Furthermore, it should be noted that the results of asymptotic analysis are merely limit results and not approximation results: they do not provide any lower bounds for $N$ such that the obtained asymptotic expressions are reasonably applicable. Consequently, asymptotic results must be used with care and should be accompanied by simulation studies (see [112] for the idea of bootstrapping for example).

This section is divided into two subsections in which the convergence and asymptotic normality of the PEM estimators defined in Section 4.3 are established. We show that the general asymptotic theory of the PEMs is applicable to the instances defined in this thesis. The results are based on the original work of Ljung in [90] and Ljung and Caines in [89] respectively. Our main focus will be the assumptions on the data, the used linear predictors and the parameterization for the results to hold.

### 4.4.1 Convergence and Consistency

Let us denote the PEM criterion function by

$$\boldsymbol{\mathcal{V}}_N(\theta) := \frac{1}{N}\sum_{t=1}^{N} \ell(\varepsilon_t(\theta), t; \theta). \tag{4.44}$$

[3] The term “asymptotic” in this thesis is restricted to large-sample scenarios, i.e., $N \to \infty$.

The PEM estimator is then defined by

$$\hat{\boldsymbol{\theta}}_N := \arg\min_{\theta\in\Theta} \boldsymbol{\mathcal{V}}_N(\theta), \quad N \in \mathbb{N}.$$

For simple cases where the PEM estimator is given in closed form (see Example 4.2.2 for instance), it is possible to establish the convergence of the estimator by directly applying variants of the law of large numbers and limit theorems concerning sequences of random variables, such as Slutsky's theorem. However, in more complicated situations where the estimator is defined implicitly with no closed-form expression, the asymptotic analysis involves the study of the asymptotic behavior of the sequence of criterion functions $\{\boldsymbol{\mathcal{V}}_N(\theta) : N \in \mathbb{N},\ \theta \in \Theta\}$ and the use of a compactness assumption on the parameter set $\Theta$ to control the corresponding process of global minimizers $\{\hat{\boldsymbol{\theta}}_N : N \in \mathbb{N},\ \theta \in \Theta\}$.

As far as the prediction error framework is concerned, the simplest cases are those involving (quasi-)stationary ergodic processes such that the sequence of criterion functions converges uniformly over $\Theta$ to a well-defined limit; namely,

$$\boldsymbol{\mathcal{V}}_N(\theta) \xrightarrow{\text{a.s.}} \bar{\mathcal{V}}(\theta) \quad \text{as} \quad N \to \infty, \tag{4.45}$$

such that the (deterministic) limit $\bar{\mathcal{V}}(\theta)$ is continuous over $\Theta$ and has a unique global minimizer $\theta^*$. In general, this limit depends on the system and the input properties. Under identifiability conditions and a compactness assumption on $\Theta$, it is straightforward to conclude, when (4.45) holds, that $\hat{\boldsymbol{\theta}}_N \xrightarrow{\text{a.s.}} \theta^* = \theta^\circ$ as $N \to \infty$. These are essentially the arguments used in the convergence and consistency proofs in an ergodic environment (see [92, Chapter 8] for the LTI case).

In a general non-stationary environment however, the sequence of criterion functions does not converge to any limit and may very well be divergent. Consider for example a linear model

$$\boldsymbol{y}_t = \theta^\circ u_{t-1} + \boldsymbol{v}_t, \quad t \in \mathbb{Z},$$

in which $\boldsymbol{v}$ is a zero mean independent process with increasing variance, and observe that the optimal MSE predictor is given by $\hat{y}_{t|t-1}(\theta) = \theta u_{t-1}$. It then holds that the average criterion function, when $\ell(\varepsilon, t; \theta) = \|\varepsilon\|^2$, is

$$\mathbb{E}\left[\boldsymbol{\mathcal{V}}_N(\theta)\right] = \sum_{t=1}^{N} \mathbb{E}\left[(\boldsymbol{y}_t - \theta u_{t-1})^2\right] = \sum_{t=1}^{N} \left((\theta^\circ - \theta)^2 u_{t-1}^2 + \mathbb{E}[\boldsymbol{v}_t^2]\right)$$

and is not necessarily convergent (even when normalized by dividing by $N$), but the global minimizer is $\arg\min_\theta \mathbb{E}\left[\boldsymbol{\mathcal{V}}_N(\theta)\right] = \theta^\circ$ $\forall N \in \mathbb{N}$. Therefore, with no (quasi-)stationary ergodic assumptions, it is possible to establish the convergence of the minimizers by showing that $\boldsymbol{\mathcal{V}}_N(\theta)$ asymptotically behaves like the averaged criterion $\mathbb{E}\left[\boldsymbol{\mathcal{V}}_N(\theta)\right]$ uniformly in $\theta$. This is the main idea of the convergence and consistency analysis developed in [90].

Before stating the basic result, we discuss sufficient regularity assumptions regarding the data, the predictor and the used criterion. As before, we assume here that the inputs are known.

### The data generating mechanism

A generic discrete-time model (2.9) of the observed process is understood as a map of sequences of known inputs and random vectors representing the underlying uncertainty. For a causal dynamical model, the output at time $t$ is a function of: (i) the known inputs up to time $t-1$ (assuming at least one delay), (ii) the disturbances entering the system after an arbitrary earlier time $k < t$, and (iii) the state of the system at time $k$ summarizing the effect of the disturbances at and before time $k$. Now observe that for an increasing sequence of observations to give a correct picture of the underlying system, the effect of the (unobserved) state at time $k$ on the outputs at times $s \geq t$ should not be predominant. In other words, the observed process should forget erroneous remote past or initial conditions, which is a property of stable systems.

For the convergence of the PE methods, it is sufficient that the dependence of the moments upon the history of the observed process decays at an exponential rate. It will be assumed that Assumption 2.2.1 holds, and therefore the terms “model” and “system” are used interchangeably.

**Definition 4.4.1** ($r$-stability). *A discrete-time causal dynamical model of $\boldsymbol{y}$ is said to be $r$-stable with some $r > 1$, if for all $s, t \in \mathbb{Z}$ such that $t \geq s$ there exist doubly-indexed random variables $\{\boldsymbol{y}_{t,s} : \boldsymbol{y}_{t,t} = 0\}$ such that*

1. *$\boldsymbol{y}_{t,s}$ is a (measurable) function of $\{\boldsymbol{y}_k\}_{k=s+1}^{t}$ and independent of $\{\boldsymbol{y}_k\}_{k=-\infty}^{s}$,*
2. *for some real numbers $C < \infty$ and $\lambda < 1$, it holds that*

$$\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_{t,s}\|^r\right] < C\lambda^{t-s}. \tag{4.46}$$

The outputs of an “$r$-stable” model according to Definition 4.4.1 form a class of stochastic processes known as “$r$-mean exponentially stable processes” or “exponentially forgetting processes of order $r$”. Observe that the definition implies that

$$\mathbb{E}\left[\|\boldsymbol{y}_t\|^r\right] < C \quad \forall t \in \mathbb{Z}, \text{ and some } C < \infty,$$

and therefore the output of an $r$-stable model must have a bounded mean. Generally speaking, the random variables $\boldsymbol{y}_{t,s}$ can be interpreted as the outputs of the system when the underlying basic stochastic process $\boldsymbol{\zeta}$ is replaced by $\{\boldsymbol{\zeta}_{t,s}\}_{t\in\mathbb{Z}}$ such that $\{\boldsymbol{\zeta}_{t,s}\}_{t<s}$ are given by a value independent of $\{\boldsymbol{\zeta}_t\}_{t<s}$, say zero, but $\boldsymbol{\zeta}_{t,s} := \boldsymbol{\zeta}_t \;\; \forall t > s$. We note here that the above definition of stability includes the conventional stability definition of dynamical systems. For example, in the case of LTI rational models, the output process is exponentially stable when all the poles of the model transfer functions are strictly inside the unit circle.

Models of Definition 2.1.8 are quite general and need to be restricted for the results to hold. Observe that not every second-order process is exponentially forgetting, even if its innovation process is independent. The next proposition clarifies this point.

**Proposition 4.4.2.** *Assume that $\boldsymbol{y}$ is a second-order discrete-time stochastic process with independent linear innovations $\{\boldsymbol{\varepsilon}_t\}$ and no linearly deterministic part; then $\boldsymbol{y}$ is not necessarily exponentially forgetting. However, if $\mathbb{E}\left[\|\boldsymbol{\varepsilon}_t\|^4\right] < \infty \;\; \forall t \in \mathbb{Z}$ and the sequences $\{h_k(t) : k \in \mathbb{N}_0\}$, $t \in \mathbb{Z}$ in Wold's decomposition (2.4) are uniformly exponentially decaying, then $\boldsymbol{y}$ is exponentially forgetting process*[4] *of order* 4.

*Proof.* The first assertion is straightforward; we only need to find an example of a second-order discrete-time stochastic process whose innovations are independent but which is not exponentially stable. Consider for example the process $\boldsymbol{y}_t = \sum_{k=1}^{\infty} k^{-1}\boldsymbol{\varepsilon}_{t-k}$ with independent innovations. This is clearly a second-order process that also forgets the remote past, however only linearly.

To prove the second part, we use Wold's decomposition of $\boldsymbol{y}$ assuming zero mean,

$$\boldsymbol{y}_t = \sum_{k=0}^{\infty} h_k(t)\boldsymbol{\varepsilon}_{t-k}, \quad t \in \mathbb{Z},$$

with the hypothesis that the innovations $\{\boldsymbol{\varepsilon}_t\}$ are independent and the sequences $\{h_k(t) : k \in \mathbb{N}_0\}$, $t \in \mathbb{Z}$ are uniformly exponentially decaying. Therefore, there exist constants $c_1 < \infty$ and $0 < \lambda < 1$ such that $|h_k(t)| < c_1\lambda^k$ for every $t \in \mathbb{Z}$.

Using the triangular inequality, it holds that for every $t \in \mathbb{Z}$ and every $n \in \mathbb{N}$

$$\begin{aligned} \left\|\sum_{k=1}^{n} h_k(t)\boldsymbol{\varepsilon}_{t-k}\right\|^4 &\leq \left(\sum_{k=1}^{N} \|h_k(t)\|\|\boldsymbol{\varepsilon}_{t-k}\|\right)^4 \leq c_1^4\left(\sum_{k=1}^{N} \lambda^k \|\boldsymbol{\varepsilon}_{t-k}\|\right)^4 \\ &\leq c_1^4\left(\sum_{k=1}^{n} \lambda^k\right)^3 \sum_{k=1}^{n} \lambda^k\|\boldsymbol{\varepsilon}_{t-k}\|^4 \end{aligned}$$

in which we used Hölder's inequality ([122, Theorem 3.5] applied to $(\lambda^{\frac{k}{p}})(\lambda^{\frac{k}{q}}\|\boldsymbol{\varepsilon}_{t-k}\|)$) for the last implication. By applying the expectation operator to both sides and letting $N \to \infty$ we get the inequality

$$\mathbb{E}\left[\|\boldsymbol{y}_t\|^4\right] \leq \frac{c_1^4}{(1-\lambda)^3}\sum_{k=1}^{\infty} \lambda^k \mathbb{E}\left[\|\boldsymbol{\varepsilon}_{t-k}\|^4\right]. \tag{4.47}$$

By defining $\boldsymbol{y}_{t,s} = \sum_{k=0}^{\infty} h_k(t)\boldsymbol{\varepsilon}_{t-k,s}$ such that $\boldsymbol{\varepsilon}_{t,s} = \boldsymbol{\varepsilon}_t$ for $t > s$ and zero otherwise, we see that (4.47) implies that

$$\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_{t,s}\|^4\right] \leq \frac{c_1^4}{(1-\lambda)^3}\sum_{k=t-s}^{\infty} \lambda^k \mathbb{E}\left[\|\boldsymbol{\varepsilon}_{t-k}\|^4\right] \leq c\lambda^{t-s}, \quad \forall t > s$$

which proves the required statement. ■

More explicit conditions can be given for specific model sets. The next example considers a Wiener model with a stochastic process disturbance.

[4] $r = 4$ is sufficient for the convergence of PEM, see Lemma 4.4.5.

**Example 4.4.1** (Exponentially stable data)**.** Consider a system that can be described by the stochastic Wiener model

$$\begin{aligned} \boldsymbol{x}_t &= G(q,\theta^\circ)u_t + H(q,\theta^\circ)w_t, \\ \boldsymbol{y}_t &= f(\boldsymbol{x}_t;\theta^\circ) + \boldsymbol{v}_t, \quad t \in \mathbb{Z}, \end{aligned} \tag{4.48}$$

and assume that the LTI part of the model is rational and stable; i.e., the poles of $G(z,\theta^\circ)$ and $H(z,\theta^\circ)$ are strictly inside the unit circle. Furthermore, assume that $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are independent and mutually independent white noises with bounded moments of sufficiently high order.

Then, in the light of Proposition 4.4.2, we see that $\boldsymbol{x}_t$ is an exponentially forgetting process. Because the nonlinearity $f(\cdot;\theta^\circ)$ is static, we only need to guarantee that several first moments of $\boldsymbol{y}$ are bounded (the order of the moments depends on $f$) and that $f(\boldsymbol{x};\theta^\circ)$ is exponentially decaying whenever $\boldsymbol{x}$ is exponentially decaying. This is the case when $f$ is a polynomial in $\boldsymbol{x}$ for example.

### The predictor

The predictor function used in the PEM framework is a user choice; it is usually guided by the assumed underlying statistical model. Apart from a natural differentiability assumption with respect to the parameter and a compactness condition on the parameter set, it is required that the remote past observation has little effect on the current output of the predictor and its derivative. From the point view of asymptotic analysis, this means that all the observed outputs, regardless of their order in time, may have a comparable contribution to the choice of the parameter. In other words, the magnitude of the prediction errors may be uniformly bounded with a suitably decaying bound. From the practical point of view, this condition is needed for the numerical stability of the minimization procedure.

This reasonable assumption means that the used predictors should have a stability property. The precise sufficient conditions are summarized in the following definition.

**Definition 4.4.3** (Uniformly stable predictors)**.** *The one-step ahead predictors* $\{\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \psi(\boldsymbol{D}_{t-1},t;\theta),\ \ \theta \in \Theta\}$ *are said to be uniformly stable if* $\Theta$ *is compact and there exist constants* $C < \infty$ *and* $\lambda \in (0,1)$ *such that the following conditions hold:*

1. $\|\xi(D_{t-1},t;\theta) - \xi(\bar{D}_{t-1},t;\theta)\| \le C\sum_{k=0}^{t-1}\lambda^{t-k}\|y_k - \bar{y}_k\|$, *where* $\xi$ *is used to denote the predictor function* $\psi$ *and it's derivative with respect to* $\theta$*, and* $D_{t-1}$*,* $\bar{D}_{t-1}$*, are arbitrary data sets of length* $t$*.*

2. $\|\xi(0,t;\theta)\| \le C\ \ \forall t,\ \forall\theta$ *in an open neighborhood of* $\Theta$*, where* $0$ *represents a data set of arbitrary inputs and zero outputs of length* $t$*.*

3. $\theta \to \psi(D_{t-1},t;\theta)$ *is continuously differentiable over an open neighborhood of* $\Theta$*, for all* $t \in \mathbb{N}$ *and for every data set* $D_{t-1}$*.*

In this thesis, we suggested two predictors: a suboptimal predictor (4.36) defined using the mean function of $\boldsymbol{y}$, and the optimal linear predictor (4.22) defined using the mean and the covariance functions $\boldsymbol{y}$. Let us examine the conditions under which they are uniformly stable according to the above definition.

First of all, note that the compactness of $\Theta$ is part of the definition. Second, it is clear that the parameterization of $\mu(U;\theta)$ and $\Sigma(U;\theta)$ is required to be continuously differentiable; this translates to smoothness conditions on the parameterization of the assumed nonlinear model. To check the remaining conditions, we first recall that the predictors have the form

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \mathbb{E}[\boldsymbol{y}_t;\theta] + \sum_{k=1}^{t-1} \tilde{l}_{t-k}(t, U_{t-1};\theta)\left(\boldsymbol{y}_k - \mathbb{E}[\boldsymbol{y}_k;\theta]\right),$$

in which

$$\tilde{l}_j(t, U_{t-1};\theta) := \left[L^{-1}(U;\theta)\right]_{tj}, \quad 1 < t \le N$$

for the optimal linear predictor, and $\tilde{l}_j(t, U_{t-1};\theta) = 0 \ \forall t$ for the OE-type linear predictor. In either case, observe that Condition 2 of Definition 4.4.3 requires that the derivative of the mean $\mathbb{E}[y_t;\theta]$ with respect to $\theta$ to be uniformly bounded in $\theta$ and $t$. We now invite the reader to compare this form to the predictors constructed for linear models in (2.39) on page 35, and recall our earlier discussion regarding the interpretation of the coefficients $\{\tilde{l}_k(t)\}$ (see Remark 4.3.5). These predictors are known to satisfy the required stability property if the transfer function $G(z;\theta)$ is stable for all $\theta \in \Theta$ in addition to the hypothesis that the noise model $H(z;\theta)$ is inversely stable (has all its zeros inside the unit circle) over $\Theta$, see [92, Lemma 4.1 and Lemma 4.2 on pages 109 and 110]. Translating this to the two predictors suggested in this thesis, we get conditions on the mean of the output process and the causal invertibility of the (innovations form) model.

Similarly to the linear case, we shall impose the assumption of causal and (exponentially) stable invertibility of $\boldsymbol{y}$ with respect to the linear innovations, and therefore the sequences $\{\tilde{l}_k(t;\theta)\}_{k\in\mathbb{N}_0}, t \in \mathbb{Z}$ are assumed to be uniformly exponentially decaying; see Assumption 2.1.7.

### The identification criterion

The identification criterion in a PEM framework (4.44) is defined by the sum of the scalar functions $\ell(\boldsymbol{\varepsilon}_t(\theta), t;\theta)$. Generally speaking, there is no unique way of defining these functions; however, a quadratic norm is the most commonly used function. It turns out that for the convergence analysis, it is sufficient to assume that these functions are quadratically bounded according to the following definition.

**Definition 4.4.4** (Quadratically bounded criteria). *The family of prediction error criterion functions $\{\boldsymbol{\mathcal{V}}_N(\theta) := \frac{1}{N}\sum_{k=1}^{N} \ell(\boldsymbol{\varepsilon}_t(\theta), t;\theta) : N \in \mathbb{N},\ \theta \in \Theta\}$ is quadratically bounded if $\{\ell(\cdot, t;\cdot)\}$ are continuously differentiable for every $t$, and for some $c < \infty$*

1. $\|\frac{\partial}{\partial\varepsilon}\ell(\varepsilon, t;\theta)\| \le c\|\varepsilon\|, \quad \forall\theta \in \Theta,$ *and* $\forall t \in \mathbb{N}$,
2. $\|\frac{\partial}{\partial\theta}\ell(\varepsilon, t;\theta)\| \le c\|\varepsilon\|^2, \quad \forall\theta \in \Theta,$ *and* $\forall t \in \mathbb{N}$.

It is clear that the unweighted Euclidean norm, $\ell(\varepsilon, t; \theta) = \varepsilon^{\mathsf{T}}\varepsilon$, used to define the L-PEM and the OE-PEM estimators is parameter-independent and quadratically bounded. For the case of WL-PEMs, the criterion is defined by using

$$\ell(\varepsilon, t; \theta) = \varepsilon^{\mathsf{T}}\Lambda_t^{-1}(\theta)\varepsilon + \log\det\Lambda_t(\theta).$$

Because this function is quadratic in $\varepsilon$, it is only required to verify that

$$\frac{\partial}{\partial\theta}\ell(\varepsilon, t; \theta) = -\varepsilon^{\mathsf{T}}\Lambda_t^{-1}(\theta)\frac{\partial\Lambda_t(\theta)}{\partial\theta}\Lambda_t^{-1}(\theta)\varepsilon + \mathbf{tr}\left(\Lambda_t^{-1}(\theta)\frac{\partial\Lambda_t(\theta)}{\partial\theta}\right)$$

is well-defined and quadratically uniformly bounded. This is a requirement on the parameterization of the covariance matrices $\Lambda_t(\theta)$ of the innovations. Observe that these matrices are defined via an $LDL^{\mathsf{T}}$ decomposition which is a continuous operation. When the parameterization is continuously differentiable such that the covariances matrices are uniformly bounded for all $t$ and $\theta$ the condition is satisfied. Therefore, once more, we end up with assumptions on the parameterization of the assumed model. We are now ready to state the basic convergence result.

**Lemma 4.4.5** (Convergence of the PEM estimators)**.** *Suppose that the nonlinear system generating the data is $r$-stable with $r = 4$, the used predictor is uniformly stable and the identification criterion is quadratically bounded. Then*

$$\hat{\boldsymbol{\theta}}_N \xrightarrow{a.s.} D_I := \left\{\theta \in \Theta : \liminf_{N\to\infty} \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)] \le \min_{\beta\in\Theta} \limsup_{N\to\infty} \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\beta)]\right\} \quad \textit{as} \quad N \to \infty.$$

*Proof.* The proof is due to Ljung; see [90]. ■

The result of Lemma 4.4.5 is very useful. Firstly, it removes the stochastic aspects of the problem and reduces the analysis to a deterministic set. Secondly, the result is proven for a fairly general case that includes scenarios where the true system is not in the assumed model set (i.e., there is no true parameter $\theta^\circ$, or $\theta^\circ \notin \Theta$). The result of the lemma means that the criterion function becomes arbitrary close to the average criterion function such that, almost surely, for every arbitrary small $\epsilon > 0$ there exist $n \in \mathbb{N}$ such that for all $N > n$, the set $D_I \cap \{\theta : \|\hat{\theta}_N - \theta\| < \epsilon\} \neq \varnothing$. Observe that, for all $\theta^* \in D_I$ it holds that

$$\liminf_{N\to\infty} \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta^*)] \le \limsup_{N\to\infty} \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)] \quad \forall\theta \in \Theta,$$

and therefore $\theta^*$ can be interpreted as a parameter that gives "the best average prediction" according to the chosen predictor and criterion function. The expectation operator here is with respect to the true underlying probability space generating the basic stochastic process.

Notice that Lemma 4.4.5 establishes the convergence of the process $\{\hat{\boldsymbol{\theta}}_N\}_{N\in\mathbb{N}}$ only to a subset of $\Theta$. However, for cases when $\theta^\circ \in \Theta$ (see Assumption 2.2.1 on page 28), and assuming that an identifiability condition holds such that the limit set is a singleton, $D_I = \{\theta^\circ\}$, a consistency proof is completed by a direct application of the lemma.

**Definition 4.4.6** (Identifiable parameterization)**.** *For a given model (2.10), we say that* $\Theta$ *constitutes*

- *a* first-order *identifiable parameterization if, for all* $\theta, \tilde{\theta} \in \Theta$*, it holds that*

$$\mu(U;\theta) = \mu(U;\tilde{\theta}) \Leftrightarrow \theta = \tilde{\theta}. \tag{4.49}$$

- *a* second-order *identifiable parameterization if, for all* $\theta, \tilde{\theta} \in \Theta$*, it holds that*

$$\mu(U;\theta) = \mu(U;\tilde{\theta}) \quad \textit{and} \quad \Sigma(U;\theta) = \Sigma(U;\tilde{\theta}) \;\; \Leftrightarrow \theta = \tilde{\theta}. \tag{4.50}$$

Note that the user is free to restrict the definition of the identifiability property to a subset of $\Theta$. For instance, the model (3.46) of Example 3.3.6 on page 68 (where $\mathbb{E}[\boldsymbol{y}_t;\theta] = u_{t-1}^2 + 2\theta^2 \;\forall t$) is not parameter identifiable over $\Theta = (-a, a)$ for any positive real $a$, but is first-order identifiable over $\Theta = (0, a)$.

**Remark 4.4.7.**

- *In the PEM literature, a martingale difference exact model for the output process is usually assumed. Under this assumption, the conditional mean of the outputs is readily available; i.e., the optimal predictor has a (directly parameterized) standard form* $\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \psi_t(\boldsymbol{D}_{t-1};\theta)$*,* $\theta \in \Theta$ *and* $t \in \mathbb{N}$*. If there exists* $\theta^\circ \in \Theta$*, the prediction error process* $\boldsymbol{e}_t(\theta^\circ) = \boldsymbol{y}_t - \hat{y}_{t|t-1}(\theta^\circ)$ *is a martingale difference and* $\mathbb{E}[\boldsymbol{e}_t(\theta^\circ)|\{\boldsymbol{e}_0(\theta^\circ), \dots, \boldsymbol{e}_{t-1}(\theta^\circ)\}] = 0$*. This is not the case for the predictors used in this thesis; for example, the prediction error process due to the optimal linear predictor (the linear innovations process) is merely orthogonal. Note that if it were a martingale difference (or independent), then the optimal linear predictor would be in fact the unrestricted optimal predictor.*
- *Assuming that the (appropriate) identifiability assumption given above holds, then the three PEM instances defined in the previous section satisfy*

$$\theta^\circ = \arg\min_{\theta\in\Theta} \;\; \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)], \qquad \forall N \in \mathbb{N}. \tag{4.51}$$

Next, we state the main consistency theorem.

**Theorem 4.4.8** (Consistency of the L-PEM and the WL-PEM estimators)**.** *Assume that the underlying nonlinear system is* $r$*-stable with* $r = 4$ *according to Definition 4.4.1, and let Assumptions 2.2.1 and 2.1.7 hold. Let* $\Theta$ *be a continuously differentiable second-order identifiable parameterization according to Definition 4.4.6 such that the optimal linear predictor is uniformly stable according to Definition 4.4.3.*

*Then, the L-PEM estimator (4.33) is consistent. Moreover, if the parameterization is such that the innovations covariances are continuously differentiable and uniformly bounded, then the WL-PEM estimator (4.35) is consistent.*

*Proof.* The proof is a direct consequence of Lemma 4.4.5 and the identifiability assumption. ■

In cases where the stronger[5] assumption of first-order identifiability holds, the consistency of the OE-PEM estimator can be established.

**Theorem 4.4.9** (Consistency of the OE-PEM estimator)**.** *Assume that the underlying nonlinear system is $r$-stable with $r = 4$ according to Definition 4.4.1, and let Assumption 2.2.1 hold. Let $\Theta$ be a continuously differentiable first-order identifiable parameterization according to Definition 4.4.6.*

*Then, the OE-PEM estimator (4.37) is consistent.*

*Proof.* The proof is a direct consequence of Lemma 4.4.5 and the identifiability assumption. ■

The last part of this section concerns the asymptotic normality of the estimators.

### 4.4.2 Asymptotic Distribution

Subject to a strengthening of the hypotheses of the consistency theorems, it is possible to prove that the resulting estimators are asymptotically normally distributed around $\theta^\circ$. We first summarize the additional required conditions without going into unnecessary details. These stronger conditions are required to establish that $\boldsymbol{\mathcal{V}}''_N(\theta)$ asymptotically behaves as $\mathbb{E}[\boldsymbol{\mathcal{V}}''_N(\theta)]$ uniformly over $\Theta$, and that the derivative $\boldsymbol{\mathcal{V}}'_N(\theta^\circ)$ is asymptotically normal when multiplied by $\sqrt{N}$ and normalized.

**Conditions for asymptotic normality**

C1. The underlying system is $r$-stable with $r > 4$ (take $r = 5$ for example; then, a uniform exponentially decaying bound on the fifth moment of $\|\boldsymbol{y}_t - \boldsymbol{y}_{t,s}\|$ is required).

C2. The used predictors are three times continuously differentiable with respect to $\theta$ such that the derivatives satisfy the first and second conditions in Definition 4.4.3.

C3. The criterion function is three times continuously differentiable with respect to $\theta$ and $\boldsymbol{\varepsilon}$ such that

1. $\|\frac{\partial^k}{\partial\theta^k}\frac{\partial}{\partial\varepsilon}\ell(\varepsilon,t;\theta)\| \leq c\|\varepsilon\|, \quad k = 0,1,2; \; \forall\theta \in \Theta, \;\text{ and }\; \forall t \in \mathbb{N},$
2. $\|\frac{\partial^k}{\partial\theta^k}\frac{\partial^2}{\partial\varepsilon^2}\ell(\varepsilon,t;\theta)\| \leq c, \quad k = 0,1; \; \forall\theta \in \Theta, \;\text{ and }\; \forall t \in \mathbb{N},$
3. $\|\frac{\partial^k}{\partial\theta^k}\ell(\varepsilon,t;\theta)\| \leq c\|\varepsilon\|^2, \quad k = 1,2,3; \; \forall\theta \in \Theta, \;\text{ and }\; \forall t \in \mathbb{N}.$

Apart from an increased smoothness requirement on the parameterization of the predictor and the criterion function, the new set of conditions require that the second and third derivatives of the predictor have the uniform stability property.

[5] Note that first-order identifiability implies second-order identifiability.

We note here that the criterion functions of the PEM instances defined in the previous section are all quadratic in $\boldsymbol{\varepsilon}$. In the case of the WL-PEM problem, the criterion is parameterized by $\theta$, and the covariances $\Lambda_t(\theta)$ have to be three times continuously differentiable and uniformly bounded according to the above requirement. This translates to a smoothness requirement on the parameterization of the covariance of the model.

**Theorem 4.4.10** (Asymptotic normality). *Assume that, in addition to the hypotheses of the consistency theorems, the set of strengthened conditions C1-C3 holds. Furthermore, let* $\mathcal{W}_N(\theta) \coloneqq \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)]$ *and assume that for some* $\delta > 0$ *and some* $N_0 \in \mathbb{N}$,

$$\mathcal{W}_N''(\theta) \succ \delta I, \quad \forall \theta \in \Theta,\ \forall N > N_0. \tag{4.52}$$

*Introduce the (normalizing) matrices*

$$P_N = [\mathcal{W}_N''(\theta^\circ)]^{-1} \mathcal{Q}_N [\mathcal{W}_N''(\theta^\circ)]^{-1},$$

*where*

$$\mathcal{Q}_N \coloneqq \mathbb{E}[N\, \boldsymbol{\mathcal{V}}_N'(\theta^\circ)(\boldsymbol{\mathcal{V}}_N'(\theta^\circ))^\top].$$

*Assume that* $P_N \succ \delta I$ *and* $\mathcal{Q}_N \succ \delta I$ *for some* $\delta > 0$ *and all sufficiently large* $N$.

*Then*

$$\sqrt{N} P_N^{-\frac{1}{2}} (\hat{\boldsymbol{\theta}}_N - \theta^\circ) \rightsquigarrow \mathcal{N}(0, I_d) \quad \textit{as} \quad N \to \infty, \tag{4.53}$$

*where* $\hat{\boldsymbol{\theta}}_N$ *denotes any of the consistent PEM estimators defined in Section 4.3.*

*Proof.* The proof is due to Ljung and Caines; see [89]. ■

Notice that the condition in (4.52) requires the strict convexity of the criterion functions $\mathcal{V}_N(\theta)$ over the whole postulated set $\Theta$ for sufficiently large $N$, which might seem quite restrictive. However, one may think of it as restricting the minimization problem to a local neighborhood of $\theta^\circ$ (using a good initial candidate $\theta^{(0)}$ in the iterative minimization algorithm).

In (quasi-)stationary ergodic scenarios where the average criterion $\mathcal{W}_N \to \bar{\mathcal{W}}(\theta)$ as $N \to \infty$ and the matrices $\mathcal{Q}_N \to \bar{\mathcal{Q}}$ as $N \to \infty$ such that the limit is invertible, it is not difficult to show that

$$\sqrt{N}(\hat{\theta}_N - \theta^\circ) \rightsquigarrow \mathcal{N}(0, P) \quad \text{as} \quad N \to \infty,$$

where

$$P = [\bar{\mathcal{W}}''(\theta^\circ)]^{-1} \bar{\mathcal{Q}} [\bar{\mathcal{W}}''(\theta^\circ)]^{-1}$$

is the asymptotic covariance matrix of the estimator. In such cases, it is actually possible to derive an expression for $P$, which is mainly used for the construction of (asymptotic) confidence intervals, the comparison between different estimators as well as experiment design. Furthermore, it is used for the design of an “optimal” criterion for a given predictor function; i.e., a criterion that leads to a minimal $P$ with respect to the usual partial ordering of positive semidefinite matrices.

With no (quasi-)stationarity assumptions or conditions, an analysis in the same spirit can be done by studying the normalizing sequence of matrices $\{P_N : N \in \mathbb{N}\}$. The scalar function $\ell$ to be preferred is that corresponding to a minimal normalizing sequence; in other words, the one leading to the largest normalization factors $P_N^{-\frac{1}{2}}$ such that the convergence in (4.53) still holds.

However, computing the expressions of $P_N$ requires the knowledge of up to the fourth moments of the innovation process for $t = 1, \dots, N$.

In Section 2.4.2, the PEMs were introduced and the kinship to the maximum likelihood method was explained. We saw that in a stochastic framework, the problems of PEMs can be interpreted as Maximum Likelihood problems. This is true even when the resulting estimator does not coincide with the true MLE of the assumed model. In the following section, we look at the Maximum Likelihood problems solved by the PEM defined in this chapter.

## 4.5 Relation to Maximum Likelihood Estimators

Let us first consider the OE-PEM estimator: the PEM based on the suboptimal linear predictor as given in Definition 4.3.10. To arrive at an equivalent maximum likelihood problem, we use the same arguments used in Section 2.4.2, page 39. It is easy to conclude that the definition of the OE-PEM estimator implicitly implies that the process $\boldsymbol{y}$ is an independent Gaussian process with unit covariance, namely

$$\boldsymbol{y}_t \sim \mathcal{N}(\mathbb{E}[\boldsymbol{y}_t;\theta], I_{d_y}),$$

and in vectors

$$\boldsymbol{Y} \sim \mathcal{N}(\mu(U;\theta), I_{d_y N}).$$

In other words, the output vector $\boldsymbol{Y}$ is jointly Gaussian with a mean vector $\mu(U;\theta)$ and a unit covariance. According to these assumptions, the likelihood function is

$$\begin{aligned}\tilde{p}(\boldsymbol{Y};\theta) &= \frac{1}{(2\pi)^{\frac{d_y N}{2}}} \exp\left(-\frac{1}{2}\|\boldsymbol{Y} - \mu(U;\theta)\|_2^2\right) \\ &= \prod_{t=1}^{N} \frac{1}{(2\pi)^{\frac{d_y}{2}}} \exp\left(-\frac{1}{2}\|\boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t;\theta]\|_2^2\right).\end{aligned}$$

and its maximization is equivalent to the problem in (4.37). It is obvious that this misspecified model captures only the first moment of $\boldsymbol{y}$. It is therefore required to assume that the model is identifiable via its first order moments.

We now look at the WL-PEM estimator: the PEM based on the optimal linear predictor as given in Definition 4.7.4. Using similar arguments as above, we see that the optimal linear predictor coincides with a conditional mean if the data follow a model

$$\boldsymbol{Y} = \mu(U;\theta) + \Sigma^{\frac{1}{2}}(U;\theta)\boldsymbol{Z} \tag{4.54}$$

in which $\boldsymbol{Z}$ is a zero mean vector with unit covariance. Under this assumption, the likelihood function is given

$$\tilde{p}(\boldsymbol{Y};\theta) = \frac{\sqrt{\det \Sigma^{-1}(U;\theta)}}{\sqrt{(2\pi)^{d_y N}}} \exp\left(-\frac{1}{2}(\boldsymbol{Y} - \mu(U;\theta))^{\top}\Sigma^{-1}(U;\theta)(\boldsymbol{Y} - \mu(U;\theta)).\right) \tag{4.55}$$

Hence, the definition of the WL-PEM estimator implicitly assumes a misspecified Gaussian likelihood function such that the vector $\boldsymbol{Y}$ has the correct mean and covariance according to the true model. It is clear that the representation in (4.54) captures the first and second moments of the model. It is therefore enough to require that the model is identifiable via its first and second moments.

**Remark 4.5.1.** *The idea of using a misspecified likelihood function to construct tractable estimators is not new. It can be traced back to [11, Section 3.3] under the name of pseudo-likelihood methods where it was used for data with spatial dependence, when the likelihood function is unavailable. It has also been suggested and studied in Econometrics; for example, the asymptotic properties were investigated in [55] for conditionally independent models.*

### Gaussian approximation of $p(\boldsymbol{Y},\boldsymbol{W};\theta)$

In this part, we would like to comment on the implicit assumption made by the PEM instances, defined in the previous section, regarding the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{W}$. It is not difficult to show that, if the distribution of $\boldsymbol{W}$ in (4.1) (the prior) is a multivariate Gaussian, then the above arguments can be extended to imply a Gaussian approximation of the joint PDF $p(\boldsymbol{Y},\boldsymbol{W};\theta)$ and the posterior $p(\boldsymbol{W}|Y;\theta)$.

As explained above, the WL-PEM instance uses the misspecified likelihood function

$$\tilde{p}(\boldsymbol{Y};\theta) = \mathcal{N}(\mu(U;\theta),\Sigma(U;\theta));$$

in addition, let us assume that

$$p(\boldsymbol{W};\theta) = \mathcal{N}(0,\Sigma_W(\theta)).$$

It is easy to conclude that (see Appendix C), for the vectors $\boldsymbol{Y}$ and $\boldsymbol{W}$ to be jointly Gaussian, it must hold that

$$\tilde{p}(\boldsymbol{Y},\boldsymbol{W};\theta) = \mathcal{N}\left(\begin{bmatrix} 0 \\ \mu(U;\theta)\end{bmatrix}, \begin{bmatrix} \Sigma_W(\theta) & \Sigma_{WY}(U;\theta) \\ \Sigma_{YW}(U,\theta) & \Sigma(U;\theta)\end{bmatrix}\right) \tag{4.56}$$

in which the covariance $\Sigma_{WY}(U,\theta)$ is defined (via the model) by

$$\Sigma_{WY}(U,\theta) := \mathbf{cov}(\boldsymbol{W},\boldsymbol{Y};\theta)$$

and $\Sigma_{YW}(U,\theta) = \Sigma_{WY}^{\top}(U,\theta)$. The approximation in (4.56) looks like the approximations suggested in Chapter 3; recall for instance Laplace's approximation in (3.23).

Even though the PEMs defined in this chapter do not require any reference to the posterior density of $\boldsymbol{W}$, (4.56) implicitly defines a Gaussian approximation. It is of interest to check their relation to the approximations of Chapter 3.

Using the standard results of conditioning Gaussian random vectors, it is easy to see that (4.56) implies the posterior

$$\tilde{p}(\boldsymbol{W}|Y;\theta) = \mathcal{N}(\mu_{W|Y}(\theta), \Sigma_{W|Y}(\theta)) \tag{4.57}$$

in which

$$\begin{aligned}\mu_{W|Y}(\theta) =& \Sigma_{WY}(U;\theta)\Sigma^{-1}(U;\theta)\left(Y - \mu(U,\theta)\right),\\ \Sigma_{W|Y}(\theta) =& \Sigma_W(\theta) - \Sigma_{WY}(U;\theta)\Sigma^{-1}(U,\theta)\Sigma_{YW}(U;\theta).\end{aligned}$$

Note that the posterior approximation (4.57), although Gaussian, is different from Laplace's approximation (3.19) which is defined by the solution of a relatively costly optimization problem over $W$. The posterior in (4.57) is defined in terms of the first two moments of the model and there is no optimization involved. While Laplace's approximation centers the Gaussian PDF at one of the modes of the posterior, (4.57) is centered at a mean value computed via Gaussian conditioning.

In case of linear Gaussian models similar to (3.33) in Example 3.3.2, both (4.57) and (3.19) are equivalent and coincide with the true posterior. However, in general, they differ and will be close only when the mode and the mean of the posterior are close. To clarify this, we repeat Example 3.3.6 where the posterior was shown to be bi-modal.

**Example 4.5.1** (Likelihood approximations of a bi-modal model)**.** Consider the model of Example 3.3.6,

$$\boldsymbol{y}_t = (u_{t-1} + \theta\boldsymbol{w}_t)^2 + \boldsymbol{v}_t, \quad t \in \mathbb{Z}, \tag{4.58}$$

in which $\boldsymbol{w}_t \sim \mathcal{N}(0,\lambda_w)$, $\boldsymbol{v}_t \sim \mathcal{N}(0,\lambda_e)$ independent over $t$ and mutually independent. The input $u_t = 0.1$ for all $t$, $\lambda_w = 2$, $\lambda_e = 0.1$, and $\theta = 0.5$. We fixed $N = 100$ and simulated one realization of the data.

Observe that the above model is an instance of (4.40), and remember that the outputs of this model are independent over time. Therefore $\Sigma(U;\theta)$ is diagonal, and it is very easy to compute the mean and the covariance of the outputs which are available in closed-form:

$$\begin{aligned}[\mu(U;\theta)]_t &= \mathbb{E}[\boldsymbol{y}_t;\theta] = u_{t-1}^2 + \theta^2\lambda_w,\\ [\Sigma(U;\theta)]_{tt} &= \mathbf{var}(\boldsymbol{y}_t;\theta) = \lambda_w\theta^2(2\lambda_w\theta^2 + u_{t-1}^2) + \lambda_v.\end{aligned}$$

Thus, the cost function of the optimal linear PEM problem (4.35) is available in closed-form and computing the WL-PEM estimate is straightforward.

Figure 4.7 shows the true negative-log likelihood and the true posterior of $\boldsymbol{w}_t$ against the approximations obtained by both Laplace's method (3.42) and the optimal linear PEM (4.35). In addition, Figure 4.8 compares the true

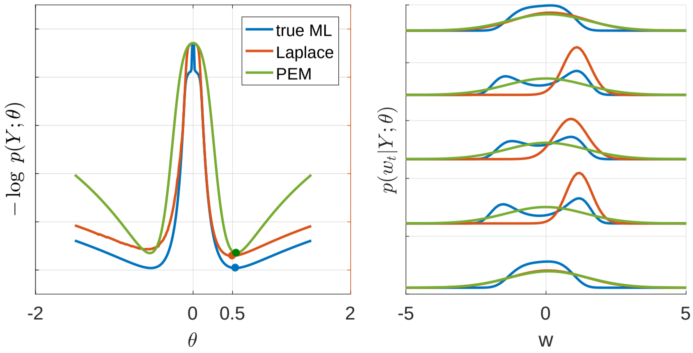


**Figure 4.7:** On the left: The true negative log-likelihood function (blue), Laplace's approximation (3.42) (red), and the cost of the optimal linear PEM (4.35) (green) of the model in (4.58). On the right: The true posterior of $\boldsymbol{w}_t$ at five selected time points (blue), Laplace's approximation (red), and the approximation underlying the optimal linear PEM (4.57) (green).

likelihood function itself to the misspecified likelihood function (4.55) underlying the optimal linear PEM.

The simulation results show that the cost functions defining the three considered estimators have the same shape, and more importantly they have very close minima. Both Laplace's approximation and PEM capture the two global minimizers of the true negative log-likelihood. However, the theoretical and computational properties of the two methods differ. While nothing can be said regarding the asymptotic properties of (3.42), the PEM estimator is consistent under mild conditions. Moreover, the PEM estimator is computationally simpler than both the true MLE and Laplace's approximation. These properties are not specific to the current example; it hold for more general models.

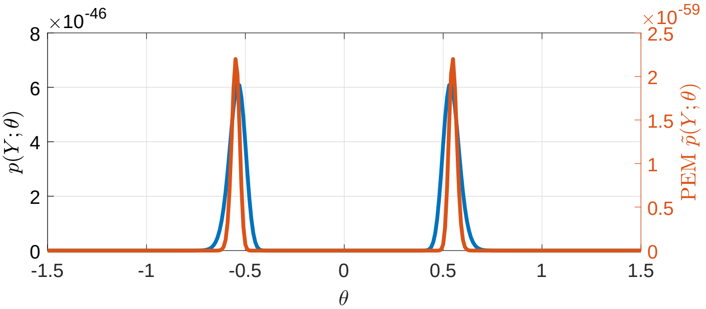


**Figure 4.8:** Plots of the true likelihood function (blue) and the misspecified likelihood (4.55) (red) of the model (4.58). Observe that regardless of the scale, both have very close global maximizers.

**Remark 4.5.2.**

- *Observe that the above ML interpretation of the PEM problem, together with the PDFs (4.56) and (4.57), which are completely defined by the given model, allows for the possibility of using the Expectation-Maximization algorithm (Algorithm 1) to solve the PEM problem. When (4.57) is used instead of Laplace's approximation, Proposition 3.3.4 on page 56 gives the expression of the intermediate quantity defining the algorithm.*

- *When using the EM algorithm to solve the PEM problem, it is still required to solve the M-step using numerical optimization. It has been claimed (see [107, Section 3.4]) that the EM iterations seem to avoid local solutions that might be problematic for a gradient-based method. However, an investigation of such a behavior has to be done; particularly in the "misspecified" setting adapted here.*

For all examples in the previous sections, it was possible to compute the linear suboptimal and linear optimal predictors in closed-form. However, for general nonlinear models, the computations of the mean and covariance of $\boldsymbol{Y}$ might be analytically intractable. In this case, it is possible to approximate them using a Monte Carlo estimator, as we show in the next section.

## 4.6 Simulated Prediction Error Method (SPEM)

To be able to solve the PEM problems (4.35), (4.33) or (4.37), it is necessary to evaluate the first two moments

$$\begin{aligned}\mu(U;\theta) &= \mathbb{E}[\boldsymbol{Y};\theta],\\ \nu(U;\theta) &= \mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^\top;\theta],\end{aligned} \tag{4.59}$$

and in turn the covariance matrix

$$\Sigma(U;\theta) = \nu(U;\theta) - \mu(U;\theta)\mu^\top(U;\theta).$$

Assume that we are given the general nonlinear model in (4.1) with one of the following assumptions:

1. The random vectors $\boldsymbol{W}$ and $\boldsymbol{V}$ follow a known joint distribution parameterized by $\theta$;
$$(\boldsymbol{W},\boldsymbol{V}) \sim p(\boldsymbol{W},\boldsymbol{V};\theta),$$
and $\boldsymbol{V}$ has a mean value $\mu_V(\theta)$.

2. The random vectors $\boldsymbol{W}$ and $\boldsymbol{V}$ are independent, $\boldsymbol{W}$ follows a known distribution parameterized by $\theta$;
$$\boldsymbol{W} \sim p(\boldsymbol{W};\theta),$$
and $\boldsymbol{V}$ has a mean value $\mu_V(\theta)$ and a covariance matrix $\Sigma_V(\theta)$.

The first case is clearly more general; however, when $\boldsymbol{V}$ is representing measurement noise, it is not restrictive to assume that it is independent of $\boldsymbol{W}$. For the more general case, the moments in (4.59) are given by the integrals

$$\mu(U;\theta) = \int \mathcal{M}(U,W;\theta)\,\mathrm{d}W + \mu_V(\theta)$$
$$\nu(U;\theta) = \iint (\mathcal{M}(U,W;\theta)+V)(\mathcal{M}(U,W;\theta)+V)^\mathsf{T} p(W,V;\theta)\,\mathrm{d}W\,\mathrm{d}V,$$

and in the case of independent $\boldsymbol{W}$ and $\boldsymbol{V}$ the second integral simplifies to

$$\nu(U;\theta) = \Sigma_V(\theta) + \iint \mathcal{M}(U,W;\theta)\mathcal{M}^\mathsf{T}(U,W;\theta)p(W;\theta)\,\mathrm{d}W.$$

Without any further assumptions (like independence of the outputs over time), these integrals are multidimensional integrals with analytic solutions only in special cases. However, even in the general situation, it is possible to estimate them using the Monte Carlo idea (see Remark 4.6.3 on the next page).

### Monte Carlo approximation of the moments

Notice that the moments of $P_\theta$ are functions of $\theta$. For every given $\theta$, we assume that it is possible to generate $M$ independent samples

$$(\boldsymbol{W}^{(m)}(\theta), \boldsymbol{V}^{(m)}(\theta)) \sim p(\boldsymbol{W},\boldsymbol{V};\theta), \quad m = 1,\ldots,M,$$

using common random numbers (see Appendix A, Section A.2). These samples together with the inputs can be used to generate pseudo realizations of the model outputs

$$Y^{(m)}(\theta) = \mathcal{M}(U, W^{(m)}(\theta);\theta) + V^{(m)}(\theta), \quad m = 1,\ldots,M.$$

Notice that the samples $Y^{(m)}$ are realization of

$$\boldsymbol{Y}^{(m)} \sim p(\boldsymbol{Y};\theta) \;\text{ independent over } m,$$

and therefore we can define the two Monte Carlo estimators

$$\begin{aligned}\widehat{\boldsymbol{\mu}(U;\theta)} &:= M^{-1}\sum_{m=1}^{M}\boldsymbol{Y}^{(m)}\\ \widehat{\boldsymbol{\Sigma}(U,\theta)} &:= (M-1)^{-1}\sum_{m=1}^{M}\left(\boldsymbol{Y}^{(m)} - \widehat{\boldsymbol{\mu}(U;\theta)}\right)\left(\boldsymbol{Y}^{(m)} - \widehat{\boldsymbol{\mu}(U;\theta)}\right)^\mathsf{T}.\end{aligned} \tag{4.60}$$

Due to the independence of the used samples, the Monte Carlo estimators in (4.60) enjoy all the desired asymptotic properties. For example, both converge almost surely to their true values as $M \to \infty$.

The idea is then to replace the intractable mean vector and covariance matrix of $\boldsymbol{Y}$ in the definition of the (sub)optimal linear predictor and the associated PEM problem by Monte Carlo estimates. We will refer to the resulting method by "Simulated PEM" (SPEM).

**Definition 4.6.1** (Simulated Prediction Error Method)**.** *The three simulated PEM estimators L-SPEM, WL-SPEM, and OE-SPEM are defined by (4.33), (4.35), and (4.37) respectively when the mean vector and the covariance matrix of the model outputs are replaced by their corresponding Monte Carlo estimates defined in (4.60).*

Notice that the resulting estimators, which we will denote by $\hat{\boldsymbol{\theta}}_M(\boldsymbol{D_N})$, are Monte Carlo approximations of $\hat{\theta}(\boldsymbol{D}_N)$ and depend on $M$. Due to the assumption that the samples $\boldsymbol{Y}^{(m)}(\theta)$ are exact independent copies over $m$, it is straightforward to conclude their convergence.

**Theorem 4.6.2** (Consistency and Asymptotic Normality of SPEM estimators)**.** *Under the same hypotheses used in Theorems 4.4.8, 4.4.9 and 4.4.10, the L-SPEM, WL-SPEM and OE-SPEM estimators are consistent and asymptotically normal if $M \to \infty$, i.e.,*

$$\lim_{M\to\infty} \hat{\boldsymbol{\theta}}_M(\boldsymbol{D_N}) \xrightarrow{a.s.} \theta^\circ \quad \text{as} \quad N \to \infty$$

*Proof.* Observe that, for a given input $u$ and a fixed $\theta \in \Theta$, the Monte Carlo estimates of the first two moments of the model are based on "exact" independent samples according to the true distribution of the outputs. Therefore, a direct use of the strong law of large numbers implies the almost sure convergence of the simulated PEM problem to the corresponding exact PEM problem. ■

**Remark 4.6.3.** *The Monte Carlo estimators of the moments suggested in this chapter are simpler than the Monte Carlo estimators of the intermediate quantity and the likelihood function from Chapter 3. Observe that here, direct sampling according to $p(\boldsymbol{W};\theta)$ is not troublesome because every sample contributes equally to the Monte Carlo sum, unlike the case of approximating marginalization integrals.*

To illustrate the idea of the SPEM, we consider the following examples.

**Example 4.6.1.** (Stochastic Wiener model with colored disturbance) Consider the following state-space model

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \theta \boldsymbol{x}_t + \boldsymbol{w}_t, \\ \boldsymbol{y}_t &= \boldsymbol{x}_t^2 + \boldsymbol{v}_t, \quad t \in \mathbb{Z} \end{aligned}$$

where $\theta = 0.7$ and $\boldsymbol{w}_t \sim \mathcal{N}(0, 0.1)$, $\boldsymbol{v}_t \sim \mathcal{N}(0, 0.1)$ both are independent of each other and independent over time. For simplicity, assume $x_0 = 0$. Observe that the model can be rewritten in the vector form

$$\boldsymbol{Y} = (\boldsymbol{X})^2 + \boldsymbol{V} = (F(\theta)\boldsymbol{W})^2 + \boldsymbol{V}$$

in which the square operator $(\cdot)^2$ is applied element-wise, and

$$F(\theta) = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \theta & 1 & 0 & \dots & 0 \\ \theta^2 & \theta & 1 & \dots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ \theta^{N-1} & \theta^{N-2} & & \dots & 1 \end{bmatrix}.$$

Even though the state process $\boldsymbol{x}$ is colored and the outputs are dependent, the model is still simple enough for analytic computations of the moments. Observe for instance that it is easy to find that

$$\mathbb{E}[\boldsymbol{Y};\theta] = 0.1(F(\theta)\mathbf{1})^2,$$

and the covariance is

$$\mathbf{cov}(\boldsymbol{Y},\boldsymbol{Y};\theta) = \mathbb{E}[(F(\theta)\boldsymbol{W})^2((F(\theta)\boldsymbol{W})^2)^\intercal] - 0.1^2(F(\theta)\mathbf{1})^2((F(\theta)\mathbf{1})^2)^\intercal$$

where once more $(\cdot)^2$ is applied element-wise. Note that it is possible to compute the second moment appearing in the covariance expression analytically; the computations require the fourth moment of $\boldsymbol{w}$ which is available. Thus, it is possible here to compute the optimal linear predictor analytically. However, for more general models, such computations can get quite involved.

In this example, we will not use analytic computations; our objective is to demonstrate the performance of the WL-SPEM estimator in a simple simulation study. Figure 4.9 shows the result of a Monte Carlo simulation over 1000 independent realizations of $\boldsymbol{w}$ and $\boldsymbol{v}$ for different values of $N$ between 100 and 1000. The number of used sample $M$ is fixed to $10^5$ for each $N$. As expected from the theory, the simulations indicate the consistency of the estimator.

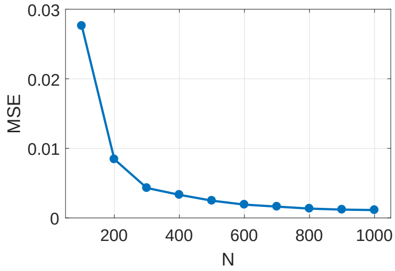


**Figure 4.9:** The average MSE over 1000 Monte Carlo simulations when the WL-SPEM is used, fixed $M = 10^5$.

In the next example, we consider a highly nonlinear model that has been used as a benchmark problem for sequential Monte Carlo methods.

**Example 4.6.2.** (The particle filter benchmark problem) Consider the nonlinear time series model

$$\begin{aligned} \boldsymbol{y}_t &= 0.05\boldsymbol{x}_t^2 + \boldsymbol{v}_t, \quad t = 1, \ldots, N, \\ \boldsymbol{x}_t &= \theta \boldsymbol{x}_{t-1} + b\frac{\boldsymbol{x}_{t-1}}{1 + \boldsymbol{x}_{t-1}^2} + 8\cos(1.2(t-1)) + \boldsymbol{w}_{t-1}, \quad \theta = 0.5, \ b = 1. \end{aligned} \tag{4.61}$$

This univariate non-stationary growth model appeared in [22, 54, 72] and has been considered by many others, see for example [21, 86, 125]. It has been mainly used as challenging benchmark problem for filtering/smoothing and parameter estimation methods due to its nonlinearity and the bi-modal posterior distribution of the state.

Observe that here, we are using a different coefficient in the second term of the state equation. The standard model has a coefficient of $b = 25$, which we replaced here by 1. The reason for this choice is that the original coefficients seem to result in abrupt and large changes in the gradient of the $\mathbb{E}[\boldsymbol{y}_t; \theta]$, see Figure 4.10.

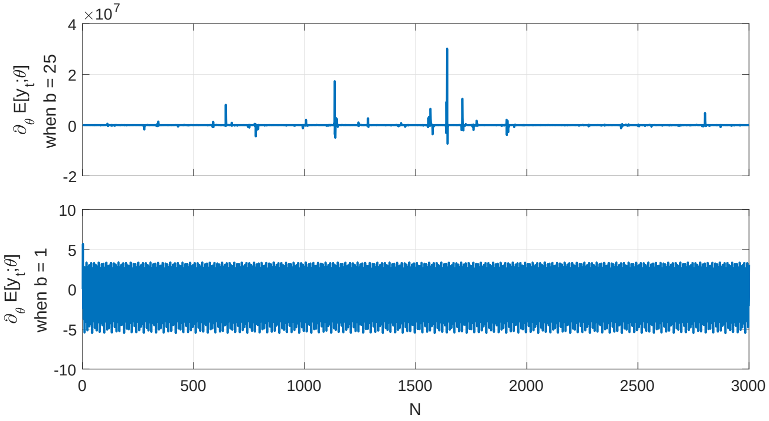


**Figure 4.10:** A sample of the derivative of the model output mean (approximated by averaging $10^5$ independent MC samples) with respect to $\theta$ when $\lambda_w = 1$, $\lambda_e = 0.1$. The case for $b = 25$ is shown in the top panel, and the case for $b = 1$ is shown at the bottom panel. In both cases, the same realization of the disturbance and noise was used.

It is obvious that the true predictor of the output and the likelihood function are both analytically intractable regardless of the distributions of $\boldsymbol{w}$ and $\boldsymbol{v}$. Observe that the measurement equation looks like the ones in Examples 4.2.2, 4.38 and 4.5.1. However, here the state process $\boldsymbol{x}$ is not white; it is a Markov process (see Definition 2.1.9) generated by $\boldsymbol{w}$ using the recursion in the second row of (4.61). The consequence is that we are not able to evaluate the moments of the outputs analytically.

To proceed, assume that $\boldsymbol{w}$ and $\boldsymbol{v}$ are independent and mutually independent Gaussian white noises with zero mean and variances $\lambda_w = \lambda_e = 1$, and let $N = 1000$. The model can be written using our vector notations in the form

$$\boldsymbol{Y} = 0.05\boldsymbol{X}^2 + \boldsymbol{V},$$

such that

$$\boldsymbol{V} \sim \mathcal{N}(0, I_N), \qquad \text{and} \qquad \boldsymbol{X} \sim p(\boldsymbol{X};\theta).$$

The PDF of $\boldsymbol{X}$ is constructed using the Markov property of the model (see (2.17));

$$p(\boldsymbol{X};\theta) = \prod_{t=1}^{N} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1};\theta).$$

Now observe that the expectation

$$\mathbb{E}[\boldsymbol{Y};\theta] = 0.05\,\mathbb{E}[\boldsymbol{X}^2;\theta]$$

is analytically intractable due to the fraction in the state equation. However, both the mean and the covariance of $\boldsymbol{Y}$ can be estimated via MC simulations.

Recall that sampling according to $p(\boldsymbol{X};\theta)$ is done sequentially as explained on pages 22 and 24 in Chapter 2. We simulated one realization of the output and solved two instances of the simulated PEM with $M = 10^4$. We assumed that the initial state $x_0 = 0$ is known. We then evaluated the cost functions for a grid of values between 0.2 and 0.8 as shown in Figure 4.11.

The results show that the WL-SPEM and OE-PEM estimators have minimizers located very close to the true value. We also observe that the cost functions in both cases have two global minima close to 0.5 and −0.5 respectively, similar to the true likelihood function (neither are shown here).

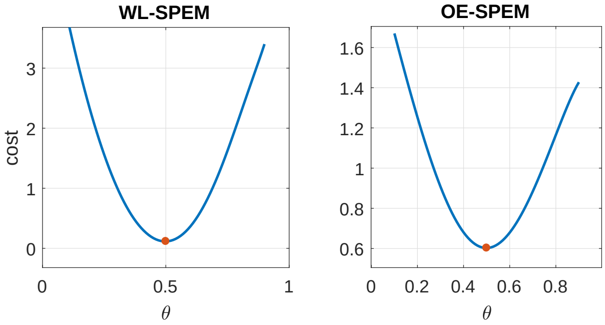


**Figure 4.11:** A sample of the cost function of $\theta$ for the WL-SPEM and the OE-SPEM assuming the model in (4.61). On the left: Monte Carlo approximation of (4.35). On the right: Monte Carlo approximation of (4.37).

Next, under the same scenario, we kept $\theta$ fixed to its true value, and evaluated the cost for different values of $\lambda_w$ between 0.25 and 4. The results are shown in Figure 4.12 where we observe a similar behavior as above.

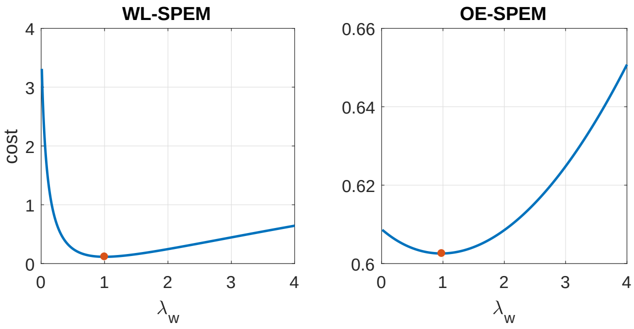


**Figure 4.12:** A sample of the cost function of $\lambda_w$ for the WL-SPEM and the OE-SPEM assuming the model in (4.61). On the left: Monte Carlo approximation of (4.35). On the right: Monte Carlo approximation of (4.37).

Table 4.1 shows the result of a Monte Carlo simulation of the WL-SPEM estimator over 1000 independent realizations of $\boldsymbol{w}$ and $\boldsymbol{v}$ and random initialization when $N = 100$. The number of used samples $M$ is fixed to $10^4$. The initial values for the parameters were chosen uniformly at random from an interval, centered at the true parameter, and have a diameter equal to 50% of the corresponding true parameter. We see that, for the assumed highly nonlinear model, the simulation results indicate the consistency of the suggested estimator.

**Table 4.1:** The mean value and the standard deviations of the WL-SPEM estimator of $\theta$ and $\lambda_w$ based on 1000 Monte Carlo runs, when $N = 100$ and $M = 10^4$.

| Parameter | True value | Estimated |
|---|---|---|
| $\theta$ | 0.5 | $0.4997 \pm 0.0223$ |
| $\lambda_w$ | 1 | $0.9883 \pm 0.208$ |

Next, we compare the WL-SPEM estimator to the state-of-the-art sequential Monte Carlo algorithms approximating the true MLE. We performed a numerical experiment similar to that in [86, Section 4], where the coefficients of the model are assumed known; but not the variances of $\boldsymbol{w}$ and $\boldsymbol{v}$. Our objective is to estimate the variances $\lambda_w$ and $\lambda_v$ in two cases: in Case 1, $\lambda_w = 1$ and $\lambda_v = 0.1$, and in Case 2, $\lambda_w = 0.1$ and $\lambda_v = 1$.

The WL-SPEM estimator (computed using an implementation of quasi-Newton algorithm) is compared to two algorithms approximating the MLE: (i) the Conditional Particle Filter with ancestor sampling used within a Stochastic Approximation Expectation-Maximization algorithm (CPF-SAEM) as suggested in [86], and (ii) the Monte Carlo Expectation-Maximization algorithm based on a (fast rejection-sampling-based) Forward Filtering/Backward SImulation (fast RS-FFBSi) smoother (also known as Particle Smoother EM (PSEM)) as suggested in [125].

For each MC realization, the three algorithms are initialized at the same random point ($\sim\mathcal{U}([1, 1.4])$ for both parameters). We used 100 MC realization with $N = 1500$, 2000 iterations for the CPF-SAEM; the step size for the stochastic approximation step is $\gamma_i = 0.98\ \forall i \leq 100$ and $i^{-0.7}$ for $100 < i \leq 2000$. Here, $M = 2 \times 10^4$ trajectories were used for the computations of the WL-SPEM estimator. For the PSEM algorithm, the number of forward filter particles is 1500 and the backward trajectories number is 300. The results of the two cases are given in Tables 4.2 and 4.3, and Figurers 4.13 and 4.14.

**Table 4.2:** The mean value and the standard deviations for the three estimators based on 100 Monte Carlo runs, when $N = 1500$, $\lambda_w = 1$ and $\lambda_v = 0.1$.

| Parameter | $\lambda_w$ (true = 1) | $\lambda_v$ (true = 0.1) | Avg. CPU time (sec.) |
|---|---|---|---|
| WL-SPEM | $0.9976 + 0.0685$ | $0.1089 + 0.0111$ | 187 |
| CPF-SAEM | $0.9982 + 0.0596$ | $0.1007 + 0.0089$ | 126 |
| PSEM | $0.9904 + 0.0595$ | $0.1024 + 0.0088$ | 7713 |

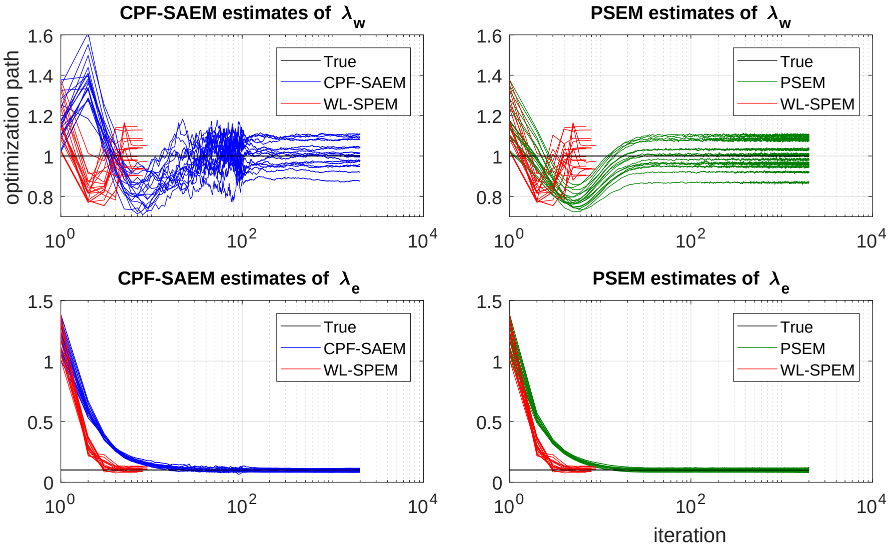


**Figure 4.13:** Parameter estimates for 20 realizations of the three estimators. Each line corresponds to one realization of the data. The true values are $\lambda_w = 1$ and $\lambda_v = 0.1$.

The simulation results show that, for the current example, the accuracy of the WL-SPEM estimator is comparable to that of the state-of-the-art MLE approximations obtained by the CPF-SAEM algorithm or the PSEM algorithm. The WL-SPEM estimator is consistent and can outperform the CPF-SAEM in computational time whenever the process disturbance variance is small or in cases where the M-step of the EM algorithm is not available in closed-form.

**Table 4.3:** The mean value and the standard deviations for the three estimators based on 100 Monte Carlo runs, when $N = 1500$, $\lambda_w = 0.1$ and $\lambda_v = 1$.

| Parameter | $\lambda_w$ (true $= 0.1$) | $\lambda_v$ (true $= 1$) | Avg. CPU time (sec.) |
|---|---|---|---|
| WL-SPEM | $0.0889 + 0.0670$ | $1.0071 + 0.0477$ | 190 |
| CPF-SAEM | $0.1871 + 0.0500$ | $0.9656 + 0.0429$ | 125 |
| PSEM | $0.0972 + 0.0604$ | $1.0029 + 0.0456$ | 17071 |

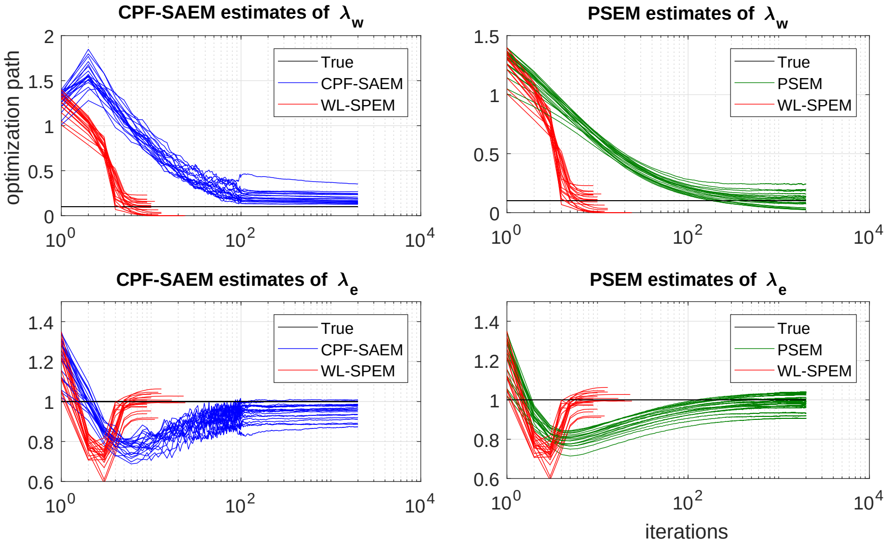


**Figure 4.14:** Parameter estimates for 20 realizations of the three estimators. Each line corresponds to one realization of the data. The true values are $\lambda_w = 0.1$ and $\lambda_v = 1$.

In the next example, we estimate two parameters of an even more challenging nonlinear state-space model of dimension 100.

**Example 4.6.3.** (High dimensional nonlinear state-space model identification) In this example, we consider a nonlinear system with 100 states and one output,

$$\begin{aligned} \boldsymbol{x}_1(t+1) &= \theta_1 \frac{\boldsymbol{x}_1(t)}{\boldsymbol{x}_{100}^2(t)+1} + \boldsymbol{w}_1(t), \qquad x_0 = 0, \\ \boldsymbol{x}_i(t+1) &= \theta_1 \frac{\boldsymbol{x}_i(t)}{\boldsymbol{x}_{i-1}^2(t)+1} + \boldsymbol{w}_i(t), \quad i = 2, \dots, 100 \\ \boldsymbol{y}_t &= \left( \sum_{i=1}^{100} x_i(k) \right)^2 + \boldsymbol{v}_t, \end{aligned} \tag{4.62}$$

in which

$$\boldsymbol{v}_t \sim \mathcal{N}(0, 0.1) \;\; \forall t, \quad \boldsymbol{w}_i(t) \sim \mathcal{N}(0, \theta_2) \;\; \forall t \text{ and } i = 1, \ldots, 100, \tag{4.63}$$

and the parameters

$$\theta_1 = 0.7, \quad \text{and} \quad \theta_2 = 0.1.$$

The state equation for each dimension is a variant of the particle filter benchmark model from the previous example. Due to its high-dimensionality, this model is quite challenging for estimators based on optimal filtering methods, especially owing to the cyclic dependence between all the states. It also poses a challenge for any estimation method that relies on approximations of the true likelihood such as the approaches in Chapter 3. To the best of the author's knowledge, the parameter identification algorithms targeting the MLE using sequential Monte Carlo algorithms have been applied only to problems with small dimensions (due to the the particle degeneracy problem).

From the WL-SPEM point of view, the problem here is similar to the first order model of the previous example and the algorithm requires no special modifications. Figure 4.15 presents the result which indicates the consistency of the estimator. Here, we estimated two parameters, one of which is the variance of the process disturbance. The first two moments of the model were approximated using $M = 10^4$ independent samples. The computational time required for one realization of the estimator is in the order of few minutes.

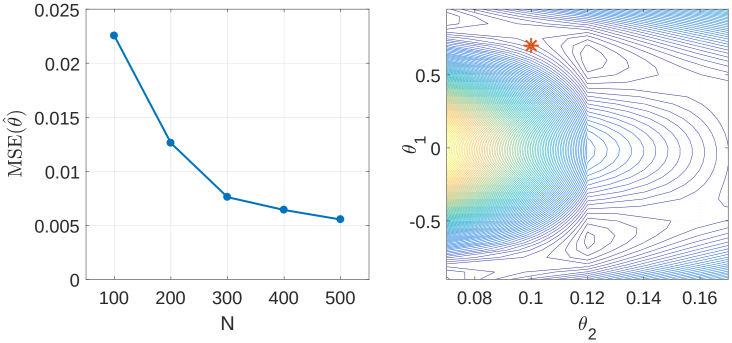


**Figure 4.15:** The results of applying the WL-SPEM estimator to the high-dimensional state-space model (4.62). On the left: The empirical MSE over 250 realizations. On the right: a sample of the contours of the cost function of the estimator. The true $\theta$ is marked with a red asterisk

---

We conclude this section with the following application-motivated example.

**Example 4.6.4** (A cascaded anaerobic digestion process)**.** Consider a dynamical model of a continuous anaerobic digestion process. This biological process is used for the treatment of organic waste in which the microorganism is broken into a mixture of methane and carbon dioxide. The details of such a process and possible models can be found in [10]. The notation in this example is slightly different from the rest of the thesis, because the model is given in continuous-time. Here, $t \in \mathbb{R}$, the subscript denotes different (dimension) variables, and time is given as an argument. It is more or less the same notation used in [10].

A stochastic two-stage bioreactor is modeled by the stochastic differential equations

$$\begin{aligned}
\mathrm{d}\boldsymbol{s}_1(t) &= [-k_1\mu_1(\boldsymbol{s}_1(t), \boldsymbol{x}_1(t)) + D_1(s_{1,\text{in}}(t) - \boldsymbol{s}_1(t))]\,\mathrm{d}t + \mathrm{d}\boldsymbol{w}_1(t)\\
\mathrm{d}\boldsymbol{x}_1(t) &= [\mu_1(\boldsymbol{s}_1(t), \boldsymbol{x}_1(t)) - D_1\boldsymbol{x}_1(t)]\,\mathrm{d}t + \mathrm{d}\boldsymbol{w}_2(t)\\
\mathrm{d}\boldsymbol{s}_{21}(t) &= [k_3\mu_1(\boldsymbol{s}_1(t), \boldsymbol{x}_1(t)) - D_1\boldsymbol{s}_{21}(t)]\,\mathrm{d}t + \mathrm{d}\boldsymbol{w}_3(t)\\
\mathrm{d}\boldsymbol{x}_2(t) &= [\mu_2(\boldsymbol{s}_{22(t)}, \boldsymbol{x}_2(t)) - D_2\boldsymbol{x}_2(t)]\,\mathrm{d}t + \mathrm{d}\boldsymbol{w}_4(t)\\
\mathrm{d}\boldsymbol{s}_{22}(t) &= [-k_2\mu_2(\boldsymbol{s}_{22}(t), \boldsymbol{x}_2(t)) + D_2(\boldsymbol{s}_{21}(t) - \boldsymbol{s}_{22}(t))]\,\mathrm{d}t + \mathrm{d}\boldsymbol{w}_5(t)
\end{aligned}$$

in which the dilution rates $D_1 = 0.04$ and $D_2 = 0.01$/day. The process $\boldsymbol{s}_1$ represents the substrate concentration in tank 1, and $\boldsymbol{s}_{1,\text{in}}$ is the substrate concentration in the influent. The processes $\boldsymbol{s}_{21}$ and $\boldsymbol{s}_{22}$ represent the product substrate concentration in tank 1 and 2 respectively, while $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are the concentrations of the biomass in tank 1 and 2 respectively. The independent Brownian motion $\boldsymbol{w}_1$ to $\boldsymbol{w}_5$ represent unobserved process disturbances. The parameters $k_1, k_2$, and $k_3$ are yield coefficients. The growth rates are modeled using a Monod-law

$$\mu_1(\boldsymbol{s}_1, \boldsymbol{x}_1) = \frac{\mu_1^* \boldsymbol{s}_1 \boldsymbol{x}_1}{\boldsymbol{s}_1 + K_{m1}}, \quad \mu_2(\boldsymbol{s}_{22}, \boldsymbol{x}_2) = \frac{\mu_2^* \boldsymbol{s}_{22} \boldsymbol{x}_2}{\boldsymbol{s}_{22} + K_{m2}},$$

with the assumption that $\mu_1^* = \mu_2^* = 1/\text{hr}$. The goal is to identify the parameter vector

$$\theta = \begin{bmatrix} k_1 & k_2 & k_3 & K_{m1} & K_{m2} \end{bmatrix}$$

using measurements of $\boldsymbol{s}_1$ and $\boldsymbol{s}_{22}$. The measurement model is given in discrete-time; the outputs are

$$\begin{aligned}
\boldsymbol{y}_1(k) &= \boldsymbol{s}_1(kT) + \boldsymbol{v}_1(k),\\
\boldsymbol{y}_2(k) &= \boldsymbol{s}_{22}(kT) + \boldsymbol{v}_2(k), \quad k = 1, \dots, N.
\end{aligned}$$

To generate the estimation data, we discretized the model using Euler's method with time interval $T = 7/24$ day. The process is then simulated in discrete-time with the initial values $\boldsymbol{s}_1 = \boldsymbol{x}_1 = \boldsymbol{s}_{22} = 0.5$, and $\boldsymbol{x}_2 = \boldsymbol{s}_{21} = 1$, and $\theta^\circ = [5\ 10\ 6\ 10\ 20]^\top$. The input $s_{1,\text{in}}$ is a square wave with levels 20 and 40 g/L. The initial level is 20 and it is changed every 7 days. The variance of the discrete-time process disturbances is $0.001T$, and the measurement noise is an independent

Gaussian process with variance 0.005. We performed a Monte Carlo experiment using 1000 realizations for three experiment durations: $60, 80,$ and 100 days. They correspond to $N = 205, 274,$ and 342 samples respectively. The number of simulations $M$ is fixed to $10^4$.

The average results of the WL-SPEM estimator are summarized in the following table

| | | $k_1$ | $k_2$ | $k_3$ | $K_{m1}$ | $K_{m2}$ | MSE |
|---|---|---|---|---|---|---|---|
| N=205 | mean | 4.99 | 9.83 | 6.01 | 9.99 | 19.6 | 8.18 |
| | std | 0.28 | 1.41 | 0.76 | 0.15 | 2.31 | |
| N=274 | mean | 5 | 9.88 | 5.99 | 9.99 | 19.7 | 5.14 |
| | std | 0.27 | 1.16 | 0.72 | 0.13 | 1.76 | |
| N=342 | mean | 4.98 | 9.93 | 6.02 | 9.99 | 19.8 | 3.81 |
| | std | 0.27 | 1.06 | 0.72 | 0.12 | 1.43 | |

As expected, the simulation results indicate the consistency of the estimator. Notice that the computational time required to estimate the parameters is, once more, in the order of a few minutes.

Before moving to the next subsection, we have the following remarks.

**Remark 4.6.4.**

- *The SPEMs are applicable to a more general class of models where* $\mathbf{V}$ *is not additive; for example, stochastic volatility models (see [20, Chapter 1]).*

- *It is possible to use a variance reduction technique (see [117, Chapter 4]) to improve the Monte Carlo estimate of the moments or to reduce the number of required samples. However, we do not pursue this possibility in this thesis.*

In the next section, we will discuss an interesting relationship between the SPEM, maximum likelihood, and the Ensemble Kalman Filter (EnKF) for the general class of nonlinear state-space model (2.11). The EnKF (see [17, 38]) is a Monte Carlo implementation of the classical Kalman filter recursions (see [70]). It was originally introduced to replace the Kalman recursions for high-dimensional problems by replacing the error covariance matrices by sample covariance matrices. Due to its flexibility, it is also used for approximate filtering in nonlinear non-Gaussian models in lieu of classical nonlinear extensions of the Kalman filter, such as [66], and particle filters (see [45], [81] and [119]). It has been used for combined state and parameter estimation in [39] but, to the best of the author's knowledge, was not considered in the PEM framework before.

## 4.7 PEM Based on the Ensemble Kalman Filter

The Ensemble Kalman Filter (EnKF) uses sequential Monte Carlo simulations to approximately compute and propagate the moments of the filtering and predictive densities of general nonlinear state-space models (2.11). It can be seen as a Monte Carlo implementation of the Kalman recursions. We start by a short presentation of the Kalman filter which will be used to motivate the EnKF.

### The Kalman Filter

Consider the LTI state-space mode used in Example 2.4.3,

$$\begin{aligned} \boldsymbol{x}_{t+1} &= A(\theta)\boldsymbol{x}_t + B(\theta)u_t + \boldsymbol{w}_t, \quad && \boldsymbol{x}_0 \sim p(\boldsymbol{x}_0;\theta), \\ \boldsymbol{y}_t &= C(\theta)\boldsymbol{x}_t + \boldsymbol{v}_t, && t \in \mathbb{N}_0, \end{aligned} \tag{4.64}$$

in which $\boldsymbol{w}$ and $\boldsymbol{v}$ are independent processes with zero mean and finite covariances $\lambda_w I_{d_w}$ and $\lambda_v I_{d_v}$. Moreover, assume that the initial state $\boldsymbol{x}_0$ has a mean value $\hat{x}_0$ and a covariance matrix $P_0$ and that $\boldsymbol{x}$ is independent of $\boldsymbol{w}$ and $\boldsymbol{v}$ for all $t$. It is well known that without any further assumptions, the Kalman filter defines the best linear estimator of the state and the best linear one-step ahead predictor of the outputs (see [2, 67]). With the additional assumption that $\boldsymbol{x}_0$, $\boldsymbol{w}$, and $\boldsymbol{v}$ are Gaussian processes, the Kalman filter gives the optimal solution to the filtering problem, i.e., the problem of finding the conditional distributions $p(\boldsymbol{x}_t|Y_t;\theta)$. Due to the linearity of the model, the filtering distributions are Gaussian in this case and they are completely defined by the mean value $\hat{x}_{t|t}$ and the covariance matrix $P_{t|t}$.

The filter proceeds recursively. Assume that the filtering density at time $t-1$ is given by

$$p(\boldsymbol{x}_{t-1}|Y_{t-1};\theta) = \mathcal{N}\left(\hat{x}_{t-1|t-1}(\theta), P_{t-1|t-1}(\theta)\right). \tag{4.65}$$

The state equation is then used to propagate $p(\boldsymbol{x}_{t-1}|Y_{t-1};\theta)$ through the dynamics of the system and the result is adjusted according to the observation $y_t$ to arrive at a filtering density at time $t$. This is done by solving the Bayesian update step

$$p(\boldsymbol{x}_t|Y_t;\theta) = \frac{p(y_t|\boldsymbol{x}_t;\theta)p(\boldsymbol{x}_t|Y_{t-1};\theta)}{p(y_t|Y_{t-1};\theta)} \tag{4.66}$$

where the prior over $\boldsymbol{x}_t$ is given by the predictive density

$$\begin{aligned} p(\boldsymbol{x}_t|Y_{t-1};\theta) &= \int p(\boldsymbol{x}_t, x_{t-1}|Y_{t-1};\theta)\,\mathrm{d}x_{t-1} \\ &= \int p(\boldsymbol{x}_t|x_{t-1}, Y_{t-1};\theta)p(x_{t-1}|Y_{t-1};\theta)\,\mathrm{d}x_{t-1} \\ &= \int p(\boldsymbol{x}_t|x_{t-1};\theta)p(x_{t-1}|Y_{t-1};\theta)\,\mathrm{d}x_{t-1} \end{aligned}$$

and the last equality follows from the Markov property (Definition 2.1.9 on page 22). Due to the linearity of the model, this integral can be solved analytically to find that

$$p(\boldsymbol{x}_t|Y_{t-1};\theta) = \mathcal{N}\left(\hat{x}_{t|t-1}(\theta), P_{t|t-1}(\theta)\right) \tag{4.67}$$

in which

$$\begin{aligned} \hat{x}_{t|t-1}(\theta) &= A(\theta)\hat{x}_{t-1|t-1}(\theta) + B(\theta)u_t \\ P_{t|t-1}(\theta) &= A(\theta)P_{t-1|t-1}(\theta)A^\top(\theta) + \lambda_w I_{d_w}. \end{aligned} \tag{4.68}$$

This is quite an elegant result. Observe how (4.65) is propagated to (4.67) by only pushing the mean and the covariance through the state equation. Using similar arguments, it holds that the one-step ahead prediction of the outputs is

$$\begin{aligned} \hat{y}_{t|t-1}(\theta) &= C(\theta)\hat{x}_{t|t-1}(\theta), \text{ with a covariance} \\ \lambda_t(\theta) &= C(\theta)P_{t-1|t-1}(\theta)C^\top(\theta) + \lambda_v I_{d_v} \end{aligned} \tag{4.69}$$

which completely defines the one-step ahead predictive density

$$p(\boldsymbol{y}_t|Y_{t-1};\theta) = \mathcal{N}\left(\hat{y}_{t|t-1}(\theta), \lambda_t(\theta)\right).$$

The likelihood of the state is easily computed using the output equation and the PDF of $\boldsymbol{v}_t$. It follows that

$$p(y_t|x_t;\theta) = \mathcal{N}\left(y_t; C(\theta)x_t, \lambda_v I_{d_v}\right).$$

Since all the PDFs are Gaussian, the filtering density at time $t$ is also Gaussian and is easily found to be

$$p(\boldsymbol{x}_t|Y_t;\theta) = \mathcal{N}\left(\hat{x}_{k|k}(\theta), P_{t|t}(\theta)\right)$$

in which

$$\begin{aligned} \hat{x}_{t|t}(\theta) &= \hat{x}_{t|t-1}(\theta) + K_t(\theta)(y_t - \hat{y}_{t|t-1}(\theta)), \\ P_{t|t}(\theta) &= P_{t|t-1}(\theta) - K_t(\theta)C(\theta)P_{t|t-1}(\theta), \text{ and} \\ K_t(\theta) &= P_{t|t-1}(\theta)C^\top(\theta)\lambda_t^{-1}(\theta) \end{aligned}$$

where the matrix $K_t(\theta)$ is known as the Kalman gain. It is used to compute the filtering density by adjusting the mean and the covariance of the predictive distribution. The steps in (4.68) and (4.69) are usually known as the time update steps, and the computations in (4.7) are known as the measurement update.

The essential step of the filter is the measurement update step. Observe that it relies on the covariance matrices of the state and the one-step ahead predictor of the outputs. The main idea of the EnKF is to avoid computing and storing the matrices $\{P_{t|t}\}$ by propagating Monte Carlo samples. Once Monte Carlo simulations are used, the method is flexible enough for general nonlinear models. In the following, we will describe the algorithm for nonlinear state-space model of the form in (2.12). We will assume that the mean value of $\boldsymbol{v}_t$ is zero for all $t$.

### The Ensemble Kalman Filter (EnKF)

To keep the notation uncluttered, the dependence of all the samples and all the estimators on $\theta$ and $M$ will not be explicit in the notation. The procedure of the EnKF is simple and intuitive. The filter starts by generating $M$ independent samples according to the prior

$$\boldsymbol{x}_0^{(m)} \sim p(\boldsymbol{x}_0;\theta), \quad \boldsymbol{w}_0^{(m)} \sim p(\boldsymbol{w};\theta) \quad \text{i.i.d. over } m = 1,\dots,M.$$

It then simulates $\boldsymbol{x}_1$ using the state equation $M$ times. Let us define the (ensemble) matrices

$$\begin{aligned} X_0 &:= \begin{bmatrix} x_0^{(1)} & x_0^{(2)} & \dots & x_0^{(M)} \end{bmatrix}, \text{ and} \\ W_0 &:= \begin{bmatrix} w_0^{(1)} & w_0^{(2)} & \dots & w_0^{(M)} \end{bmatrix}. \end{aligned}$$

with as many rows as the state dimension and $M$ columns. Then, the predictive ensemble at time $t = 1$

$$\begin{aligned} X_{1|0} &= \begin{bmatrix} x_{1|0}^{(1)} & x_{1|0}^{(2)} & \dots & x_{1|0}^{(M)} \end{bmatrix} \\ &= h(X_0, u_0 \mathbf{1}^\top, W_1; \theta) \end{aligned}$$

where the symbol $h$, with a slight abuse of notation, is used to also operate on ensembles (element-wise). The symbol $\mathbf{1}$ denotes a column vector of ones and has a dimension $M$. Generally, at time $t > 1$, assume that an ensemble $X_{t-1|t-1}$ is available. The filter simulates the time updates

$$\begin{aligned} X_{t|t-1} &= \begin{bmatrix} x_{t|t-1}^{(1)} & x_{t|t-1}^{(2)} & \dots & x_{t|t-1}^{(M)} \end{bmatrix} \\ &= h(X_{t-1|t-1}, u_{t-1} \mathbf{1}^\top, W_{t-1}; \theta) \end{aligned} \tag{4.70}$$

in which

$$W_{t-1} = \begin{bmatrix} w_{t-1}^{(1)} & w_{t-1}^{(2)} & \dots & w_{t-1}^{(M)} \end{bmatrix}$$

and $\boldsymbol{w}_{t-1}^{(m)} \sim p(\boldsymbol{w};\theta)$ i.i.d. over $t$ and $m$. The next step is to simulate the (predicted) output as follows

$$\begin{aligned} Y_{t|t-1} &= \begin{bmatrix} y_{t|t-1}^{(1)} & y_{t|t-1}^{(2)} & \dots & y_{t|t-1}^{(M)} \end{bmatrix} \\ &= g(X_{t|t-1};\theta) + V_t. \end{aligned} \tag{4.71}$$

in which

$$V_t = \begin{bmatrix} v_t^{(1)} & v_t^{(2)} & \dots & v_t^{(M)} \end{bmatrix}$$

and $\boldsymbol{v}_t^{(m)} \sim p(\boldsymbol{v};\theta)$ i.i.d. over $t$ and $m$. Observe that, with a slight abuse of notation, the symbol $g$ is used to also operate on ensembles.

The two steps in (4.70) and (4.71) are analogous to the time update step of the Kalman filter. Observe that the samples (ensemble or particles) $X_{t|t-1}$ and $Y_{t|t-1}$ provide empirical approximations to the predictive densities $p(\boldsymbol{x}_t|Y_{t-1};\theta)$ and

$p(\boldsymbol{y}_t|Y_{t-1};\theta)$ respectively. Similarly, $X_{t-1|t-1}$ is seen as an empirical approximation of the filtering density $p(\boldsymbol{x}_{t-1}|Y_{t-1};\theta)$. The important observation here is that whenever $X_{k-1|k-1}$ contains exact i.i.d. samples according to the filtering density $p(\boldsymbol{x}_{t-1}|Y_{t-1};\theta)$, the samples of $X_{t|t-1}$ and $Y_{t|t-1}$ will be exact i.i.d. samples of the respective distributions. This observation suggests the use of the following Monte Carlo estimators of the mean

$$\begin{aligned}\hat{x}_{t|t-1} \approx \tilde{x}_{t|t-1} &:= M^{-1} X_{t|t-1}\mathbf{1},\\ \hat{y}_{t|t-1} \approx \tilde{y}_{t|t-1} &:= M^{-1} Y_{t|t-1}\mathbf{1},\end{aligned} \tag{4.72}$$

and the covariances

$$\begin{aligned}\mathbf{cov}(\boldsymbol{x}_{t|t-1},\boldsymbol{x}_{t|t-1}) &\approx P_{t|t-1} = (M-1)^{-1}\,\tilde{X}_{t|t-1}\tilde{X}_{t|t-1}^{\top},\\ \mathbf{cov}(\boldsymbol{y}_{t|t-1},\boldsymbol{y}_{t|t-1}) &\approx \lambda_t = (M-1)^{-1}\,\tilde{Y}_{t|t-1}\tilde{Y}_{t|t-1}^{\top} + \lambda_v I_{d_v},\\ \mathbf{cov}(\boldsymbol{x}_{t|t-1},\boldsymbol{y}_{t|t-1}) &\approx S_t = (M-1)^{-1}\,\tilde{X}_{t|t-1}\tilde{Y}_{t|t-1}^{\top},\end{aligned} \tag{4.73}$$

where

$$\begin{aligned}\tilde{X}_{t|t-1} &= X_{t|t-1} - \tilde{x}_{t|t-1}\mathbf{1}^{\top}, \text{ and}\\ \tilde{Y}_{t|t-1} &= g(X_{t|t-1};\theta) - \tilde{y}_{t|t-1}\mathbf{1}^{\top}.\end{aligned}$$

If the ensemble $X_{t-1|t-1}$ were exact with i.i.d. samples, a standard version of the law of large numbers would apply and the above mean and covariance estimators would converge almost surely to their true values as $M \to \infty$. However the used samples are dependent and the way the filter employs the observations and the prior at $t$ to compute the ensemble $X_{t-1|t-1}$ relies on couple of approximations as we shall now explain.

The ensemble $X_{t|t}$ is computed based on an approximate solution of the Bayesian update (4.66). The joint predictive density $p(\boldsymbol{x}_t,\boldsymbol{y}_t|Y_{t-1};\theta)$ is approximated by the multivariate Gaussian distribution

$$p(\boldsymbol{x}_t,\boldsymbol{y}_t|Y_{t-1};\theta) \approx \mathcal{N}\left(\begin{bmatrix}\tilde{x}_{t|t-1}\\ \tilde{y}_{t|t-1}\end{bmatrix}, \begin{bmatrix}P_{t|t-1} & S_t\\ S_t^{\top} & \lambda_t\end{bmatrix}\right),$$

where the means and covariances were defined above in (4.72) and (4.73). By conditioning on $\boldsymbol{y}_t$, we get a measurement update equation similar to the linear case; namely

$$\check{\boldsymbol{x}} = \hat{x}_{t|t-1}(\theta) + K_t(\theta)(y_t - \check{\boldsymbol{y}}),$$

in which $\check{\boldsymbol{x}}$ and $\check{\boldsymbol{y}}$ are place holders. This relation defines the mechanism used by the EnKF to update each element of the ensembles, that is

$$X_{t|t} = X_{t|t-1} + \mathcal{K}(y_t\mathbf{1}^{\top} - Y_{t|t-1}) \tag{4.74}$$

in which the effect of $\mathcal{K}$ is defined by multiplying each column of the argument matrix by the gain

$$K_k = S_k\lambda_k^{-1} = \frac{1}{M-1}\sum_{1=m}^{M} \tilde{X}_{t|t-1}\tilde{Y}_{t|t-1}^{T}\lambda_k^{-1}. \tag{4.75}$$

This is analogous to the Kalman gain in the linear case. We note here that in practice, the empirical covariance matrix $P_{t|t}$ does not require to be computed or stored. The computations of the Kalman gain $K_k$ are sample-based, which allows the filter to work with very high-dimensional state-space models. In addition, because the time update step is based on simulations, the filter can be applied to any nonlinear model that can be simulated.

In the following example, we compute the log-likelihood function of a linear Gaussian model by using the EnKF.

**Example 4.7.1.** This example demonstrates the convergence of the EnKF in the linear case. We consider a model similar to the one in (2.31) in Example (2.4.1);

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \theta \boldsymbol{x}_t + \boldsymbol{w}_t, \\ \boldsymbol{y}_t &= \boldsymbol{x}_t + \boldsymbol{v}_t, \quad t \in \mathbb{N}_0, \end{aligned} \tag{4.76}$$

such that $\theta = 0.7$, $\boldsymbol{w}$ and $\boldsymbol{v}$ are stationary independent and mutually independent Gaussian white noises with unit variance. With $N = M = 1000$, we simulated one realization of the model outputs and evaluated the negative log-likelihood function at several values of $\theta$. The results in Figure 4.16 show that the EnKF (almost) coincides with log-likelihood as computed by the sequential KF or direct evaluations.

We note here that the convergence of the EnKF to the KF for linear models is not a trivial result. Observe that the used samples in (4.72) and (4.73) are not independent. Each element of the filtering ensemble depends on the whole prediction ensemble $X_{t|t-1}$, and therefore the samples are dependent; however, the dependence is only through the empirical covariance matrix computations. A law of large numbers still holds as shown in [81] where the convergence of the EnKF is established.

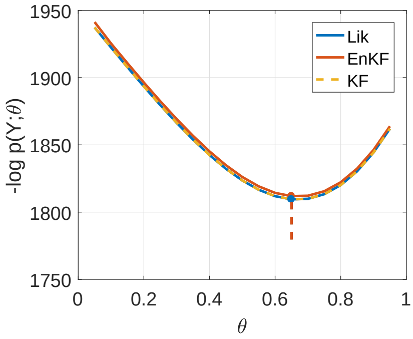


**Figure 4.16:** Comparison between EnKF and KF for the LTI model in (4.76). The figure shows the negative log-likelihood computed with three methods: (i) Direct computation (in blue), (ii) sequential computation using the EnKF (in red), and (iii) sequential computation using the time-varying KF (in yellow)

### PEM based on the EnKF

In this last part, we first define a PEM instance based on the one-step ahead predictors (4.72) given by the EnKF. We then investigate the relationship between the resulting PEM estimator and the SPEM estimator based on the suboptimal and the optimal linear predictors we defined earlier in this chapter.

**Definition 4.7.1** (The EnKF one step-ahead predictors)**.** *The EnKF one-step ahead predictors of the outputs at time $t$ are defined by*

$$\tilde{y}_{t|t-1}(\theta) := \frac{1}{M} Y_{t|t-1}(\theta)\mathbf{1} \tag{4.77}$$

*in which $Y_{t|t-1}$ is the predictive ensemble of the outputs.*

**Proposition 4.7.2.** *The EnKF one-step ahead predictor of $\boldsymbol{y}_t$ is nonlinear in both previous outputs $\boldsymbol{Y}_{t-1}$ and previous inputs $U_{t-1}$.*

*Proof.* To see this, it is only required to write the predictors in terms of the inputs and the outputs. Note that the predictive ensembles are defined by propagating the ensemble members through the nonlinear state and output equations, see (4.70) and (4.71). Therefore, assuming that $\boldsymbol{v}_t$ has zero mean, the predictors are given by

$$\begin{aligned}\tilde{y}_{t|t-1}(\theta) &= \frac{1}{M}\sum_{m=1}^{M} g(x^{(m)}_{t|t-1};\theta)\\ &= \frac{1}{M}\sum_{m=1}^{M} g(h(x^{(m)}_{t-1|t-1}, u_{t-1}, w^{(m)}_{t-1};\theta);\theta)\end{aligned}$$

The state filtering ensemble is updated according to (4.74), and we end up with

$$\tilde{\boldsymbol{y}}_{t|t-1}(\theta) = \frac{1}{M}\sum_{m=1}^{M} g\Bigg( h\Big(\overbrace{ \underbrace{x^{(m)}_{t-1|t-2} + K_{t-1}(\boldsymbol{y}_{t-1} - y^{(m)}_{t|t-1})}_{x^{(m)}_{t-1|t-1}}, u_{t-1}, w^{(m)}_{t-1} ;\theta\Big)}^{x^{(m)}_{t|t-1}} ;\theta\Bigg)$$

■

It is obvious that the EnKF one-step ahead predictors are different from the suboptimal and optimal linear predictors defined earlier this chapter. However, as the following lemma asserts, there are some cases where the EnKF is related to the OE-type suboptimal predictor.

**Lemma 4.7.3.** *Assume that the state process and the output process of the state-space model are such that $\mathbf{cov}(\boldsymbol{x}_t, \boldsymbol{y}_t) = 0$ for all $t \in \mathbb{Z}$. Then the EnKF one-step ahead predictor $\tilde{y}_{t|t-1}(\theta)$ is a Monte Carlo approximation of $\mathbb{E}[\boldsymbol{y}_t;\theta]$ and therefore is measurement independent and coincides with the suboptimal predictor defined in (4.60).*

*Proof.* The proof is straightforward. In all cases where $\mathbf{cov}(\boldsymbol{x}_t, \boldsymbol{y}_t) = 0$, the Kalman gain is $K_t = 0$, see (4.75). In this case, the measurement update step has no effect and the measurement is never used to update the prediction ensemble. This means that for all $t$, it holds that $X_{t|t} = X_{t|t-1}$ and therefore the samples of the ensemble $Y_{t|t}$ are just i.i.d. samples distributed according to $p(\boldsymbol{y}_t; \theta)$. Consequently, the predictor (4.77) is equivalent to the first row of (4.60). ■

**Example 4.7.2.** Consider the case of a stochastic Wiener model

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \theta \boldsymbol{x}_t + \boldsymbol{w}_t \\ \boldsymbol{y}_t &= \boldsymbol{x}_t^2 + \boldsymbol{v}_t, \quad t \in \mathbb{N}_0 \end{aligned}$$

where $\boldsymbol{x}_0 \sim \mathcal{N}(0, \lambda_{x_0})$, $\boldsymbol{w}_t \sim \mathcal{N}(0, \lambda_w)$ and $\boldsymbol{v}_t \sim \mathcal{N}(0, \lambda_v)$ for all $t$. Furthermore, assume that $\boldsymbol{v}_t$ is independent of both $\boldsymbol{x}_0$ and $\boldsymbol{w}_t$ for all $t$. Then, $\boldsymbol{x}_t$ is a Gaussian random variable and $\mathbf{cov}(\boldsymbol{x}_t, \boldsymbol{y}_t) = 0$ for every $t$. Thus, the predictor (4.77) is a Monte Carlo approximation of the suboptimal linear predictor (4.36) (it does not depend on $\boldsymbol{y}_t$).

In the following definition, a PEM estimator is defined based on the EnKF one-step ahead predictors.

**Definition 4.7.4** (PEM based on the EnKF)**.** *The PEM estimator based on the EnKF is defined by*

$$\begin{aligned} \hat{\theta}_M(\boldsymbol{D}_N) = \arg\min_{\theta \in \Theta} \quad & \|\boldsymbol{Y} - \tilde{\boldsymbol{Y}}(\theta)\|^2_{\Lambda^{-1}(U;\theta)} + \log\det \Lambda^{-1}(U;\theta) \\ \textit{such that} \quad & \tilde{\boldsymbol{Y}}(\theta) := \begin{bmatrix} \tilde{\boldsymbol{y}}^\top_{1|0}(\theta) & \tilde{\boldsymbol{y}}^\top_{2|1}(\theta) & \cdots & \tilde{\boldsymbol{y}}^\top_{N|N-1}(\theta) \end{bmatrix}^\top, \end{aligned}$$

*where the one-step ahead linear predictions $\tilde{\boldsymbol{y}}_{t|t-1}$ are defined in Definition 4.7.1, and $\Lambda(U;\theta)$ is a diagonal matrix with entries $\lambda_t(\theta)$ as defined in (4.73).*

It is clear that this estimator looks very similar to the estimator based on the simulated optimal linear predictor as defined in (4.6.1). Both solve a prediction error problem with parameterized norm using simulated predictors (also note that an unweighted version may also be defined). However, they differ in the way they define the predictor and the covariances. The SPEM estimators defined in (4.6.1) is based on a Monte Carlo approximation of "linear" predictors. However, as shown above, the EnKF predictors are nonlinear and it is not obvious how to assess the properties of the resulting estimator.

To get an idea about the relation between the two, we introduce the following example.

**Example 4.7.3.** First, we consider a model for which the relationship between $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ is linear.

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \theta \frac{\boldsymbol{x}_t}{\boldsymbol{x}_t^2 + 1} + \boldsymbol{w}_t, \qquad && \boldsymbol{w}_t \sim \mathcal{N}(0, 0.1) \\ \boldsymbol{y}_t &= \boldsymbol{x}_t + \boldsymbol{v}_t, && \boldsymbol{v}_t \sim \mathcal{N}(0, 0.1), \end{aligned} \tag{4.78}$$

We assume that $x_0 = 0$, $\theta = 0.7$ and $N = 100$. We then simulate one realization of the outputs and compute the one-step ahead predictors (4.77) and a simulated version of the optimal linear predictor (4.22) using the true $\theta$ and $M = 10^5$. Figure 4.17 shows plots of the prediction errors, prediction error variance, and the one-step ahead predictions, and the cost function for both estimators for a grid of values for $\theta$. To further control the comparison, we use the same random numbers for both cases.

Although not identical, the EnKF predictor follows the optimal linear predictor closely. More interestingly, the cost functions have the same shape and seem to have very close minimizers.

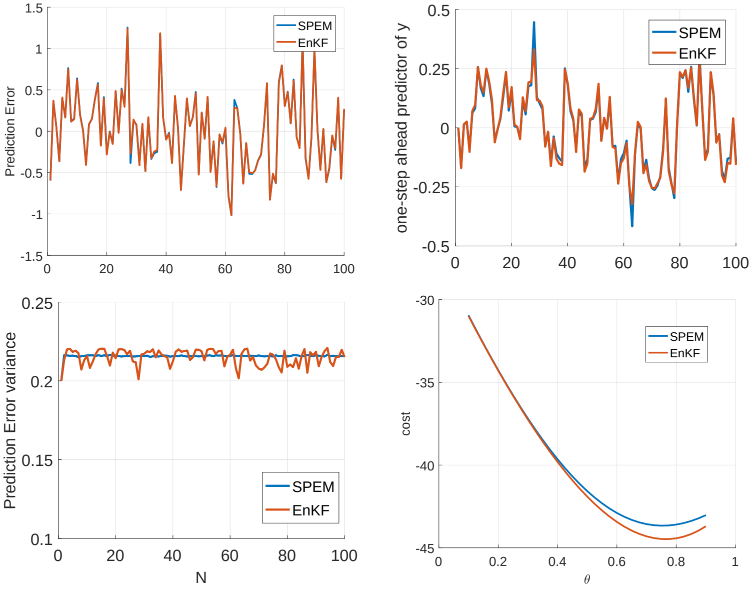


**Figure 4.17:** Simulation results for the model in (4.78).

Next, we repeat the same experiment with the same setting but using a quadratic observation model

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \theta \frac{x_t}{x_t^2 + 1} + \boldsymbol{w}_t, \\ \boldsymbol{y}_t &= \boldsymbol{x}_t^2 + \boldsymbol{v}_t. \end{aligned} \tag{4.79}$$

Figure 4.18 shows similar conclusions to those found in the case of a linear observation model. This experiment seems to highlight a close relationship between the two estimators.

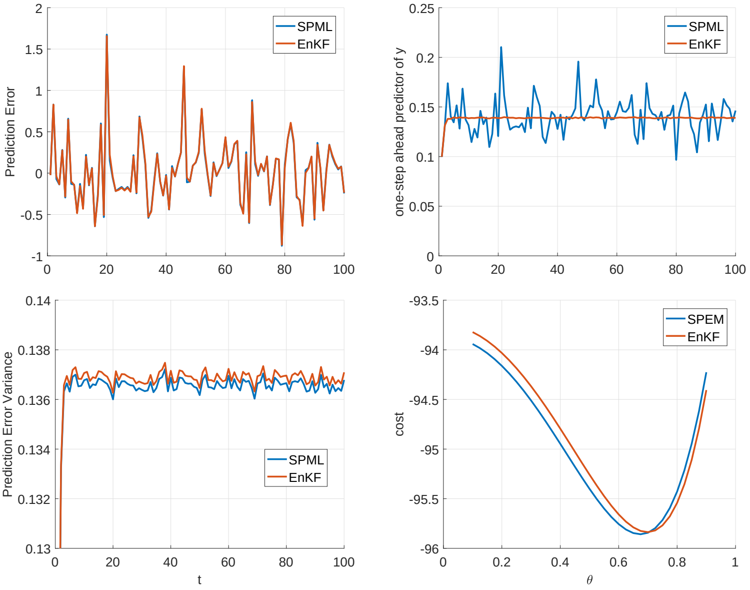


**Figure 4.18:** Simulation results for the model in (4.79).

## 4.8 Summary

In this chapter, we introduced a consistent and asymptotically normal estimator for parametric nonlinear models. As explained in Section 4.2, the basic idea is to apply the PEM using predictors that are linear in the observed output, but may be nonlinear in the known input. The given examples show that in several cases of interest, the predictors and the objective functions are given in closed-form. The optimal linear (in $\boldsymbol{y}_t$) one-step ahead predictor and the associated PEM problem are defined in Section 4.3. Several remarks were given regarding the interpretation of the predictor, the connection to Wold's decomposition and the linear case. The asymptotic analysis of the suggested estimators is given in Section 4.4. We discussed the required conditions on the data, the predictors, and the used criterion and parameterization for the general results in [90] and [89] to hold. A maximum likelihood interpretation is given in Section 4.5. We then defined the simulated PEM: a Monte Carlo approximation of the PEM estimators suggested in this thesis. They are used when the first two moments of the model are analytically intractable. Finally, we discussed the Ensemble Kalman Filter (EnKF) and an associated PEM. We saw that the EnKF one-step ahead predictor is nonlinear in both $\boldsymbol{y}_t$ and $u_t$, however in some cases it might be independent of the measurement.

Chapter 5

# Conclusions and Future Research Directions

The content of this thesis concerns the estimation problem of parametric nonlinear stochastic dynamical models. The problem is considered under the following assumptions:

- The model structure is known,
- There exists an unobserved (latent) stochastic process influencing the outputs through a non-invertible nonlinear relation.
- The data is generated in open-loop, i.e., the input signal is known and is independent of all other signals.

In this setting, the commonly used point estimation methods such as the maximum likelihood method and the prediction error method –relying on the optimal one-step ahead predictor– are analytically intractable. While ignoring the existence of the unobserved stochastic disturbance leads to tractable problems, it is well known that the resulting estimators are biased.

Recently, there has been extensive research in the system identification community on the use of sequential Monte Carlo and/or Markov Chain Monte Carlo methods to solve general nonlinear inference problems. These methods have been shown to provide acceptable solutions on several academic examples; nevertheless, scaling these methods to high-dimensional models with many parameters is quite challenging and is currently an active research area. Moreover, the convergence of some of these methods is guaranteed only within the exponential family.

In an attempt to address these problems, the thesis is divided into two main parts:

1. approaches for approximate solutions of the maximum likelihood problem, and
2. consistent instances of the prediction error methods.

In the following sections, we first summarize the conclusions of the thesis and then give several pointers for possible future research.

## 5.1 Thesis Conclusions

Let us begin by commenting on Wold's decomposition given in Chapter 2.

### Wold's decomposition

Theorem 2.1.6 gives an interesting insight regarding the structure of general second-order non-stationary processes. It shows that any purely non-deterministic process with finite second-order moments can be seen as causally and linearly filtered white noise (innovations). Two observations are important here: (i) the filters are time-varying and not necessarily exponentially stable (the impulse response sequences may only be square summable), (ii) the innovations are merely uncorrelated and may be dependent.

In linear system identification, the basic stochastic process is assumed stationary, "linea" and purely non-deterministic, i.e., it is defined as the output of a stable LTI filter whose input, the innovations sequence, is an i.i.d. process. These stronger assumptions imply the summability of the covariance sequence, and therefore it is possible to define a spectral density for the process. However, observe that the linearity and stability of the process do not imply the (causal and stable) invertibility with respect to the innovations. Therefore, further conditions have to be "imposed" to guarantee invertibility. The spectral factorization theorem gives sufficient conditions on the power spectrum that can be translated to properties of the used transfer operator or state-space models.

For the nonlinear models considered in this thesis, Wold's decomposition is used to motivate the use of linear predictors. The invertibility assumption – similar to the linear case – is imposed. However, direct verifiable conditions on the underlying model were not studied.

### Chapter 3

In this chapter, we explored several analytical and numerical approximations for the maximum likelihood estimator.

### The EM and the quasi-Newton algorithms

The EM algorithm is advantageous whenever the E-step and the M-step are tractable with closed-form expressions. In this situation, the algorithm does not require the computation of any gradients. However, this advantageous situation is restricted to only some models; e.g., joint models in the exponential family. On the other hand, the quasi-Newton algorithm is a gradient-based algorithm and therefore it is parameterization-dependent. Nevertheless, due to the use of Hessian approximations, the quasi-Newton algorithm can be a few times faster compared to the EM. Both algorithms are intractable for general nonlinear models. The EM requires approximations of the posterior PDF of the disturbance, and the quasi-Newton algorithm needs approximations of the (log-)likelihood function and its gradient.

### Analytic approximations of the EM algorithm

The convergence results of the EM algorithm rely mainly on the fact that the posterior distribution used to define the $Q$-function corresponds exactly to the integrated joint model. When the posterior is replaced by any other non-degenerate distribution, the convergence of the algorithm is not guaranteed (even if the integral w.r.t the approximate PDF is evaluated exactly). This is what we observed when Laplace's approximation of the posterior was used to integrate the true joint model.

On the other hand, assuming that the covariance of Laplace's approximation is infinite such that the PDF is concentrated on the MAP estimate of the disturbance, the EM algorithm is equivalent to a joint estimation problem. The resulting estimator is known to be biased (even in the linear case), nevertheless it has been also observed that it could be preferable to the MLE for short data records.

### Analytical approximations of the likelihood function

The normalization constant of Laplace's approximation of the posterior can be used as an approximation of the log-likelihood function. Regardless of the posterior distribution, the quality of the log-likelihood approximation relies on how accurately a multivariate Gaussian can approximate the posterior around any of its possibly many modes. The resulting estimator can be seen as a regularized joint MAP estimator that coincides with the MLE in the linear case. The simulation examples indicate that the approximation might be acceptable in some cases.

One disadvantage common to all methods based on analytic approximations is the difficulty of analyzing the resulting estimates. At best, we were only able to show that these approximations are exact for the linear case.

### Numerical Approximations

Whenever the dimension of the disturbance process is small enough and the outputs are independent over time, deterministic numerical integration can be efficiently used to approximate the iterates of the EM algorithm. To do so, the interesting observation that the $Q$-function can be written in terms of an integral with respect to the (assumed known) PDF is used. For models with colored outputs, the involved integrals are multidimensional and deterministic integration methods are hopeless. In such cases, MC approximations should be used instead.

The $Q$-function and the likelihood function can be written as expectations with respect to the known distribution of the latent process. Thus, direct sampling may be used to define an unbiased estimator of each. However, due to the nature of the conditional likelihood, the variance is usually very large for any manageable number of samples. In this thesis, we used Laplace's approximation of the posterior as an importance sampling density. In principle, this leads to a variance reduction when $N$ is small and the true posterior is relatively close to a Gaussian. However, for larger values of $N$, the required number of samples might be very large.

## Chapter 4

In this chapter, we looked at several consistent instances of the PEM.

### Linear predictors and PEMs for nonlinear models

Motivated by Wold's decomposition of the outputs, we proposed linear predictors based on the underlying nonlinear model. The predictors are linear in the observed outputs; however, the dependence on the known inputs may be nonlinear.

The simplest possible predictor of any model is given by the mean value of its outputs. In cases where the output process is independent, the mean value is in fact the unrestricted optimal predictor (it coincides with the conditional mean). Using this simple predictor, we defined what we call an OE-type (Output-Error type) PEM. As the name suggests, this is equivalent to the use of an OE-structure in linear system identification.

If the first two moments of the model are available, the optimal linear predictor can be derived. Similarly to the linear case, the optimal linear predictor is constructed by inverting a noise model. However, unlike the linear case, the noise model here is time-varying because it depends on the used input. Under the assumption of zero initial conditions, only the first part of the noise model impulse response is required. Based on the optimal linear predictor, what we call an L-PEM (Linear PEM) can be defined. In the linear case, this is just the commonly used PEM formulation.

The accuracy of the OE-PEM and L-PEM estimators can be improved by the use of weighting. The given examples provided several possible alternatives. The best accuracy is obtained when the L-PEM problem is weighted using the "time-varying" covariances of the innovations. Because these are unknown and depend on the parameter, a $\log\det$ term needs to be added to the cost function to establish the consistency of the estimator. We denote the resulting estimator by WL-PEM (Weighted Linear PEM).

In cases where the first two moments of the model are analytically intractable, we proposed the use of vanilla MC approximations. Here, direct sampling is not troublesome because every sample contributes equally to the Monte Carlo sum, unlike the case of approximating marginalization integrals. The obtained approximations can be used in lieu of the exact moments to define the OE-SPEM (Output-Error Simulated PEM), L-SPEM (Linear Simulated PEM) and WL-SPEM (Weighted Linear Simulated PEM) estimators which converge to the corresponding exact versions as the number of MC samples approaches infinity.

The asymptotic theory of the PEMs is applicable to the instances defined in this thesis. Under some mild conditions on the data, model parameterization, and the mean of the model output, the proposed estimators are all consistent and asymptotically normal. Several simulation examples, including some challenging models, confirm these results.

### Maximum likelihood interpretations

The OE-PEM, L-PEM, and WL-PEM problems have an interesting interpretation in terms of misspecified Maximum Likelihood problems. They all correspond to a misspecified multivariate Gaussian model for the model outputs. All the instances correctly specify the mean of true model, however they differ in the specification of the covariance (the weighting). This interesting point of view makes it possible to obtain a Gaussian approximation for the joint model (of the outputs and the unobserved disturbance) whenever the disturbance process is Gaussian. This approximation, while in the same spirit, is different from Laplace's approximation. The joint Gaussian approximation allows for the possibility of using the EM algorithm to solve the PEM problems.

### The EnKF and a corresponding PEM

The EnKF is an MC version of the basic KF which can be used for nonlinear state-space models. Similar to the simulated linear predictors proposed in this thesis, the EnKF predictor depends on the first two moments of the model. However, it is shown that the obtained predictor is nonlinear in both the inputs and the outputs. Nevertheless, it boils down to the mean of the outputs whenever the covariance of the state and the output is zero. When the covariance is not exactly zero but small, the EnKF resembles the OE-type predictor. However, no asymptotic guarantees can be made in general.

## 5.2 Possible Future Research Directions

In this last section, we give some ideas for possible future research.

### Asymptotically efficient two-step estimator

The PEM estimators developed in this thesis are computationally attractive. In several relevant cases, they may be defined using closed-form predictors and, under some mild conditions, they are consistent and asymptotically normal, but not asymptotically efficient. It is in fact possible to improve the asymptotic properties of these consistent estimators by just one iteration of a Newton-Raphson scheme. Denote the consistent PEM estimator by $\hat{\boldsymbol{\theta}}$; then, it can be shown that the estimator

$$\tilde{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}} - [\nabla_\theta^2 \log p(Y, W; \hat{\boldsymbol{\theta}})]^{-1} \nabla_\theta \log p(Y, W; \hat{\boldsymbol{\theta}})$$

is asymptotically efficient. The symbols $\nabla_\theta p(Y, W; \hat{\theta})$ and $\nabla_\theta^2 \log p(Y, W; \hat{\theta})$ denote the gradient of the log-likelihood function and its Hessian respectively, evaluated at $\hat{\theta}$. This means that an efficient estimator can be constructed by a "single" evaluation of the gradient and the Hessian of the log-likelihood function. This can be achieved by solving (3.47) and a similar integral for the Hessian using only a 'single' run of any Monte Carlo smoothing algorithm, for example a conditional particle filter. The resulting two-step procedure is expected to be computationally cheaper compared to a full SMC solution.

### Initialization and relaxation methods for Wiener-Hammerstein models

One disadvantage of the (S)PEMs in general is the difficulty of the involved optimization problems. According to the assumed model structure, the resulting optimization problem is usually non-convex with several local minima, especially with short data records. Thus, a good initial guess of the parameters is needed.

For Wiener-Hammersten models, it is well known that under some assumptions on the noise, the Best Linear Approximation is directly related to the LTI part of the model. One possibility here is the use of weighted least-square methods to obtain an initialization point.

### Using the Expectation-Maximization algorithm to solve SPEM

In this thesis, we proposed the SPEMs and gave them an ML interpretation. This opened the possibility of approximating the joint density $p(\boldsymbol{Y}, \boldsymbol{W}; \theta)$ by a Gaussian PDF whenever $p(\boldsymbol{W}; \theta)$ is Gaussian. An interesting possibility would be to use the expectation-maximization algorithm to solve the SPEM problem. In particular, a Stochastic Approximation EM might be advantageous for several reasons: (i) i.i.d. exact samples can be easily generated, (ii) the number of used samples does not need to grow with $N$, (iii) a convergence guarantee is available (see [28]), (iv) it has been observed that SAEM prevents the sequence from staying near an unstable stationary point of the likelihood and might behave better than the EM algorithm in some cases (see [105]). Therefore, it is of interest to determine whether a stochastic approximation EM algorithm can be used.

### Simulation studies and comparisons

In this thesis, the simulation and numerical examples were chosen to clarify the ideas and expose the advantages and disadvantages of the proposed methods. For example, a main goal was to verify the consistency of the methods in various situations. However, a simulation study comparing the performance of the proposed SPEM to that of the available alternative is required for better understanding of the methods.

### Variance reduction techniques for SPEM

Even though the variance of the Monte Carlo estimators of the first two moments does not depend on $N$, it depends on the variance of the model outputs. If this variance is large, vanilla Monte Carlo methods might be inefficient. In this case, it is possible to adopt variance reduction techniques.

### Nonlinear errors-in-variables estimation problems

The methods proposed in this thesis were developed and analyzed under the assumption that the input is known. In some practical situations, the user might not be in control of the identification experiment, and therefore both inputs and outputs have to be measured. Several challenging technical problems arise in this situation.

Appendix A

# The Monte Carlo Method

In this appendix, we describe the Monte Carlo idea for approximating PDFs and integrals. For more details, we refer the reader to any of the several books on the Monte Carlo methods. See for example, the books [117], [88], or [37].

## A.1 The Monte Carlo Idea

The Monte Carlo (MC) idea is based on replacing some unknown or intractable probability distribution function by an empirical distribution based on a set of random samples.

Consider a random variable $\boldsymbol{\zeta}$ defined on some set $\mathsf{Z}$ (usually a Euclidian real space) and distributed according to a probability distribution function $F(\boldsymbol{\zeta})$; in addition, consider a sequence of independent random variables

$$\boldsymbol{\zeta}^{(m)} \sim F(\zeta), \quad m = 1, \dots, M$$

that are copies of $\boldsymbol{\zeta}$. These random variables can be used to define the empirical distribution

$$\mathrm{d}\boldsymbol{F}_M \coloneqq \frac{1}{M} \sum_{m=1}^{M} \delta_{\boldsymbol{\zeta}^{(m)}}(\mathrm{d}\boldsymbol{\zeta}), \tag{A.1}$$

in which $\delta_{\boldsymbol{\zeta}^{(m)}}(\mathrm{d}\boldsymbol{\zeta})$ denotes a Dirac measure on the singleton $\{\boldsymbol{\zeta}^{(m)}\}$, and a corresponding probability density function

$$\hat{\boldsymbol{p}}_M(\boldsymbol{\zeta}) \coloneqq \frac{1}{M} \sum_{m=1}^{M} \delta_{\boldsymbol{\zeta}^{(m)}}(\{\boldsymbol{\zeta}\}).$$

Let $\varphi$ be some test function, defined over $\mathsf{Z}$, such that it is integrable with respect to $F$, and consider the problem of evaluating the integral

$$\mathbb{E}[\varphi(\boldsymbol{\zeta})] \coloneqq \int_{\mathsf{Z}} \varphi(\zeta)\, \mathrm{d}F(\zeta).$$

A MC estimator of this integral can be defined using (A.1) as the “random variable”

$$\hat{\mathbb{E}}_M[\varphi(\boldsymbol{\zeta})] := \int_{\mathsf{Z}} \varphi(\zeta)\,\mathrm{d}\boldsymbol{F}_M = \frac{1}{M}\sum_{m=1}^{M}\varphi(\boldsymbol{\zeta}^{(m)}). \tag{A.2}$$

A MC estimate is given by a realization $\{\zeta^{(m)}\}_{m=1}^{M}$ and is written as a MC sum

$$\hat{\mathbb{E}}_M\,[\varphi(\boldsymbol{\zeta})] = \int_{\mathsf{Z}} \varphi(\zeta)\,\mathrm{d}F_M = \frac{1}{M}\sum_{m=1}^{M}\varphi(\zeta^{(m)}).$$

It is immediate that the estimator (A.2) of $\mathbb{E}[\varphi]$ is unbiased and, due to the assumption that the random variables $\boldsymbol{\zeta}^{(m)}$ are independent over $m$, a direct application of the strong law of large numbers (see [23, Chapter 5]) shows that

$$\hat{\mathbb{E}}_M[\varphi(\boldsymbol{\zeta})] \xrightarrow{\text{a.s.}} \mathbb{E}[\varphi(\boldsymbol{\zeta})] \quad \text{as} \quad M \to \infty. \tag{A.3}$$

The symbol $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence. This means that, if we use sufficiently large number of samples $M$, we can achieve any required approximation accuracy of the expectation of $\varphi$ by averaging the values $\{\varphi(\zeta^{(m)})\}$. It is also not difficult to see that when the function $\varphi$ has a finite variance (with respect to $F$), a standard version of the central limit theorem (see [23, Chapter 7]) implies the convergence in distribution of the normalized (MC) errors; that is

$$\sqrt{N}\left(\hat{\mathbb{E}}_M[\varphi(\boldsymbol{\zeta})] - \mathbb{E}[\varphi(\boldsymbol{\zeta})]\right) \rightsquigarrow \mathcal{N}(0, \mathbf{var}(\varphi)) \quad \text{as} \quad M \to \infty \tag{A.4}$$

in which the symbol $\rightsquigarrow$ denotes convergence in distribution. This last result can be used to construct asymptotic confidence regions for the estimator in (A.2). The important observation to be made here is that the variance of the estimator does not depend directly on the dimension $N$. This is a notable advantage of the Monte Carlo method over deterministic approximation methods. The accuracy of the Monte Carlo method depends however on the number of used samples $M$ and the variance $\mathbf{var}(\varphi)$ of the integrand under the true measure. It follows that the approximation error decreases at a rate of $\mathcal{O}(M^{-1})$; however it should be noticed that the proportionality constant can be quite large depending on how the samples are generated.

Monte Carlo estimators rely on the assumption that it is possible to generate i.i.d. samples according to $F$. However, this is usually not possible in practice. Even if it is possible to generate i.i.d. samples, and depending on the properties of the integrand function, the required number of samples $M$ to achieve a certain accuracy might be prohibitively large. Therefore, most of the research performed on methods relying on the MC idea try to answer one or both of the following two questions:

1. How to sample according to a high-dimensional distribution that has an intractable probability density/distribution function?

2. How to decrease the computational complexity and accelerate the convergence of the method in the sense of minimizing the number of required samples $M$?

In the following sections, we explain briefly what is meant by generating a random sample according to some distribution.

## A.2 Random Sampling and Common Random Numbers

The MC idea presented in the previous section is based on the assumption that it is possible to generate (using a computer machine) as many realizations as wished of certain random variables. In some cases, it is required to do so for a random variable whose PDF has no explicitly known closed-form expression.

We first consider the problem of producing realizations of a uniform random variable in the interval $[0,1]$. The first difficulty that one faces when trying to solve such a problem is how to deal with the philosophical notion of randomness. Without dwelling on such a notion, what we really want to obtain is a completely deterministic method known as a "uniform pseudo-number generator". This is a well defined algorithm characterized by a transformation $\mathcal{T}$ on the unit interval. The transformation defines a recursion, that when started at a known initial (deterministic) value $\zeta^{(0)} \in [0,1]$, called the "seed", it produces a sequence of values

$$\{\zeta^{(m)}\} = \{\mathcal{T}^m(\zeta^{(0)})\} \subset [0,1], \quad 1 \leq m \leq M$$

such that the statistical hypothesis

$$H_0: \quad \boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)}, \dots, \boldsymbol{\zeta}^{(M)} \text{ are i.i.d.} \sim \mathcal{U}([0,1])$$

is accepted under a usual family of uniformity and independence tests, see [83, Chaper 14] for example. This means that the elements of this sequence are required to behave statistically in a similar way as i.i.d. samples of $\mathcal{U}([0,1])$. In such a case, we allow ourselves to write

$$\boldsymbol{\zeta}^{(m)} \sim \mathcal{U}([0,1]).$$

Due to the deterministic nature of the algorithm, using the same seed $\zeta^{(0)}$ will always produce the same unique sequence. Many of the available software packages like MATLAB, Mathematica, Julia, R, ..., etc. come equipped with efficient pseudo-random number generators. The user is able to control the seed of the algorithm every time a random sample is generated, which makes all MC simulations repeatable. In this case, we say that the MC method is using "common random numbers". Using common random numbers is required for the algorithms developed in this thesis (to preserve the continuity and smoothness of the objective functions, see [49]).

Because any generic probability space $(\Omega, \mathcal{F}, P)$, can be constructed by defining random variables $\boldsymbol{\zeta}$ over the basic probability space $([0,1], \mathcal{B}([0,1]), \mathcal{U}([0,1]))$ such that, $\boldsymbol{\zeta} : [0,1] \to \Omega$ (see [73, Theorem 1.104] and recall that $\mathbb{R}$ is equipotent to the unit interval), it is evident that a uniform pseudo-number generator is sufficient to produce pseudo-realizations of many random variable (at least theoretically). This is indeed the case if the distribution function $F$ is known. Assume that we have $\boldsymbol{\zeta} \sim \mathcal{U}([0,1])$, then it is not difficult to show that $F^{-1}(\boldsymbol{\zeta})$ is a pseudo-random sample according to the given distribution $F$ (see [88, Lemma 2.1.1]). In practice, however, this approach can be used only when $F$ is available in closed-form and its inverse can be easily evaluated. Even if these two conditions hold, generating samples this way might not be the best option in terms of algorithmic efficiency. We refer the interested reader to [30, 74, 116].

In the next section, we present the importance sampling method that can be used as a variance reduction technique or in situations where direct sampling is not possible, or is very time-consuming.

## A.3 Importance Sampling

The importance sampling solution can be traced back to the beginning of MC techniques. It has been introduced in [59] and [121] as a variance reduction technique for MC approximation methods. It has also been used in [44] for Bayesian inference and in [50] for the simulation of Markov chains. The importance sampling idea is nothing more than a change of measure trick. Assume that we are interested in generating a random sample according to a probability density function $p$, but we have one of the following situations:

1. the normalizing constant of $p$ is unknown or sampling from $p$ is difficult or time-consuming.
2. $p$ is known and easy to sample from, but the resulting MC estimators have high variance.

In either case, we may introduce an "importance sampling density" $q$ (also known as proposal density) defined on the same probability space, and for any value $\zeta$ we define the ratio

$$w(\zeta) = \frac{p(\zeta)}{q(\zeta)}$$

which is known in this context as the "importance weight" or merely the weight. For these values to be finite for all $\zeta$, we require that the support of $q$, i.e., the set

$$\operatorname{supp}(q) = \{\zeta \mid q(\zeta) \neq 0\},$$

contains the support of $p$. Because both densities are defined on the same space, the random samples $\{\boldsymbol{\zeta}^{(m)} \sim q,\ m = 1, \dots, M\}$ with the weights $\{w(\boldsymbol{\zeta}^{(m)}) : m = 1, \dots, M)\}$ can be used to define the empirical probability density function $\hat{\boldsymbol{p}}_M$,

$$\hat{\boldsymbol{p}}_M(\boldsymbol{\zeta}) = \frac{1}{M} \sum_{m=1}^{M} w(\boldsymbol{\zeta}^{(m)})\ \delta_{\boldsymbol{\zeta}^{(m)}}(\{\boldsymbol{\zeta}\}).$$

See Figure A.1 for an example of importance sampling of a triangular distribution.

When the random variables $\boldsymbol{\zeta}^{(m)}$ are independent over $m$ and $p$ and $q$ are exactly known, the strong law of large numbers implies that

$$\frac{1}{M} \sum_{m=1}^{M} \boldsymbol{w}_m \xrightarrow{\text{a.s.}} 1 \quad \text{as} \quad M \to \infty \tag{A.5}$$

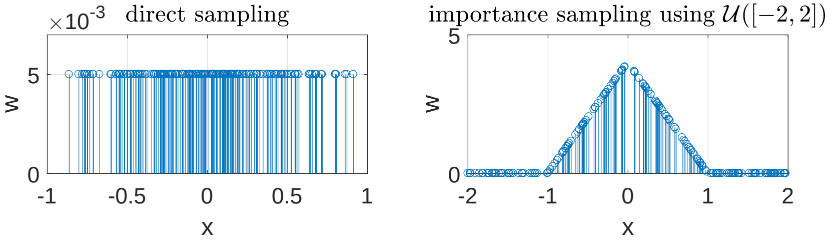


**Figure A.1:** Importance sampling of a triangular distribution with a PDF $p(x) = x+1$ whenever $-1 \leq x \leq 0$, $p(x) = 1 - x$ whenever $0 < x \leq 1$, and $p(x) = 0$ whenever $|x| > 1$. The plot on the right shows 200 direct samples, and the plot on the left shows 200 weighted sampled generated according to $\mathcal{U}([-2,2])$.

in which

$$\boldsymbol{w}_m := w(\boldsymbol{\zeta}^{(m)}) = \frac{p(\boldsymbol{\zeta}^{(m)})}{q(\boldsymbol{\zeta}^{(m)})}.$$

The estimator of $\int \varphi(\zeta)p(\zeta)\,\mathrm{d}\zeta$, for any integrable $\varphi$, given by

$$\hat{\mathbb{E}}_M[\varphi(\boldsymbol{\zeta})] = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{w}_m \varphi(\boldsymbol{\zeta}^{(m)}), \text{ in which } \boldsymbol{\zeta}^{(m)} \sim q. \tag{A.6}$$

is unbiased for any value $M$, and almost surely consistent. The central limit theorem for i.i.d. samples can be directly applied and deviation inequalities might be used to check the accuracy of the estimates.

### Self-normalized importance sampling

In many situations, the target distribution $p$ is known only up to a normalization factor. This means that the importance weight function is known only up to a constant scaling factor, and the convergence (A.5) does not hold anymore. In this case, importance sampling can still be used by adopting the self-normalized form. For this, normalized importance weights (self-normalized weights) are defined by

$$\bar{\boldsymbol{w}}_m := \frac{\tilde{\boldsymbol{w}}_m}{\sum_{m=1}^{M} \tilde{\boldsymbol{w}}_m},$$

in which

$$\tilde{\boldsymbol{w}}_m = \frac{\tilde{p}(\boldsymbol{\zeta}^{(m)})}{\tilde{q}(\boldsymbol{\zeta}^{(m)})}$$

and

$$\tilde{p}(\boldsymbol{\zeta}^{(m)}) \propto p(\boldsymbol{\zeta}^{(m)}), \quad \tilde{q}(\boldsymbol{\zeta}^{(m)}) \propto q(\boldsymbol{\zeta}^{(m)}).$$

Then, the self-normalized form of the empirical distribution is given by

$$\hat{\boldsymbol{p}}_M(\boldsymbol{\zeta}) = \sum_{m=1}^{M} \bar{\boldsymbol{w}}_m\, \delta_{\boldsymbol{\zeta}^{(m)}}(\{\boldsymbol{\zeta}\}) = \frac{\frac{1}{M}\sum_{m=1}^{M} \tilde{\boldsymbol{w}}_m\, \delta_{\boldsymbol{\zeta}^{(m)}}(\{\boldsymbol{\zeta}\})}{\left(\frac{1}{M}\sum_{m=1}^{M} \tilde{\boldsymbol{w}}_m\right)}.$$

This is a ratio of two sample means that, by the strong law of large numbers, converges almost surely to $p$, but is biased for any finite value $M$. The resulting weighted sample $\{(\zeta^{(m)}, \bar{w}_m)\}$ is said to be consistent for $p$.

The importance sampling idea is quite general; it introduces only little restrictions on the choice of the proposal density $q$ which is assumed to be easy to sample from. The idea can even be taken further by assuming that the proposal density itself is given in terms of a weighted sample $\{(\zeta^{(m)}, \bar{w}_q^{(m)})\}$, where $\bar{w}_q^{(m)}$ denotes the $m^{\text{th}}$ weight with respect to the PDF $q$. In this case, the importance sampling algorithm makes the transformation

$$\{(\zeta^{(m)}, \bar{w}_q^{(m)})\} \mapsto \{(\zeta^{(m)}, \bar{w}_p^{(m)})\}$$

by only modifying the weights. More complex transformations can be applied to modify both the samples and the weights.

One interesting property of the importance sampling method is that the consistency (of $\hat{\mathbb{E}}_M\left[\varphi(\boldsymbol{\zeta})\right]$ or $\hat{\boldsymbol{p}}(\boldsymbol{\zeta})$) can be established in some cases where the MC samples are dependent. For instance, as shown in [96], it is possible to use MCMC samplers within an importance sampling algorithm without introducing any bias. However, this generality and flexibility of the method hides the difficulty of the original problem in the step of choosing a proposal density $q$. Even though the method officially makes use of all the samples, a careless choice for $q$ might lead to very small weights that are practically 0. In this case, most of the samples will not contribute to the MC sum and the method will be inefficient. In other words, only very few samples (in the worst case, only one) will contribute to the approximation of the target distribution.

Appendix B

# Hilbert Spaces of Random Variables

In this appendix, we review some relevant definitions, properties and theorems of the Hilbert space of random variables with zero mean and finite second moment. The proofs of all the statements and more details can be found in any book on functional analysis. See for example [122], [151], [115], or [3].

## B.1 Inner Product Spaces

Let $\mathcal{H}$ be a vector space over the reals $\mathbb{R}$. One way to define a topological structure over $\mathcal{H}$ is by defining an inner product, denoted $\langle \cdot, \cdot \rangle$. The pair $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is known as an inner product space or a pre-Hilbert space.

**Definition B.1.1** (Inner product space)**.** *A real vector space $\mathcal{H}$ is an inner product space if there is an inner product (a function)*

$$\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_+,$$

*that is a mapping such that for any three vectors $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z} \in \mathcal{H}$ and two scalars $\alpha, \beta \in \mathbb{R}_+$,*

*i.* $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$,

*ii.* $\langle \alpha \boldsymbol{x} + \beta \boldsymbol{y}, \boldsymbol{z} \rangle = \alpha \langle \boldsymbol{x}, \boldsymbol{z} \rangle + \beta \langle \boldsymbol{x}, \boldsymbol{y} \rangle$,

*iii.* $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0 \quad \forall\, \boldsymbol{x} \in \mathcal{H}$ *and* $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0 \iff \boldsymbol{x} = 0$.

An inner product can be used to define a norm.

**Definition B.1.2** (Induced norm)**.** *The inner product of an inner product space $\mathcal{H}$ can be used to define the quantity*

$$\|\boldsymbol{x}\|_{\mathcal{H}} = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} \text{ for any vector } \boldsymbol{x} \in \mathcal{H}.$$

*$\|\cdot\|_{\mathcal{H}}$ is said to be the induced norm on $\mathcal{H}$.*

It is easy to check that the induced norm is indeed a norm on $\mathcal{H}$, and therefore $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ is a normed space. It can be used to define usual topological concepts over $\mathcal{H}$, such as closure, openness, convergence and completeness. On the other hand, the underlying inner product can be used to define geometrical concepts like orthogonality, parallelism and angles between vectors. This allows us to generalize the intuitive geometrical concepts from the standard Euclidean space $\mathbb{R}^3$ to abstract infinite dimensional function spaces.

**Definition B.1.3** (Orthogonality). *Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be an inner product space. Any two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{H}$ are said to be orthogonal if*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0.$$

*This is symbolized by writing $\boldsymbol{x} \perp \boldsymbol{y}$. A vector $\boldsymbol{x} \in \mathcal{H}$ is said to be orthogonal to a set $S$ (written $x \perp S$) if $x \perp s$ for all $s \in S$.*

**Definition B.1.4** (Cauchy Sequence). *A sequence $\{\boldsymbol{x}_n : n = 1, 2, \dots\}$ of vectors in an inner-product space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ is said to be a Cauchy sequence if*

$$\|\boldsymbol{x}_n - \boldsymbol{x}_m\|_{\mathcal{H}} \to 0 \text{ as } m, n \to \infty.$$

**Definition B.1.5** (Hilbert space). *A Hilbert space is a complete inner product space. That is, every Cauchy sequence converges in the topology generated by the induced norm.*

**Definition B.1.6** (Closed subspaces). *A linear subspace $\mathcal{S}$ of $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ is said to be a closed subspace if for any sequence $\{\boldsymbol{x}_n\} \subset \mathcal{S}$ and some $\boldsymbol{x} \in \mathcal{H}$*

$$\|\boldsymbol{x}_n - \boldsymbol{x}\|_{\mathcal{H}} \to 0 \text{ as } n \to \infty \implies \boldsymbol{x} \in \mathcal{S}.$$

**Lemma B.1.7** (Finite dimensional subspaces). *Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a Hilbert space. Any finite dimensional subspace $S$ of $\mathcal{H}$ is closed in $\mathcal{H}$.*

**Definition B.1.8** (Orthogonal complement). *For any $\mathcal{S} \subset \mathcal{H}$, its orthogonal complement is*

$$\mathcal{S}^{\perp} := \{\boldsymbol{x} \in \mathcal{H} : \langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0 \ \ \forall \boldsymbol{y} \in \mathcal{S}\}.$$

An important theorem of Hilbert spaces is the projection theorem.

**Theorem B.1.9** (The projection theorem). *If $\mathcal{S}$ is a closed subspace of $\mathcal{H}$ and $\boldsymbol{x} \in \mathcal{H}$, then:*

*i. There is a unique element $\hat{\boldsymbol{x}} \in \mathcal{S}$ such that*

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_{\mathcal{H}} = \inf_{y \in \mathcal{S}} \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{H}}.$$

*ii. If $\hat{\boldsymbol{x}} \in \mathcal{S}$ then $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_{\mathcal{H}} = \inf_{y \in \mathcal{S}} \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{H}}$ if and only if $(\boldsymbol{x} - \hat{\boldsymbol{x}}) \in \mathcal{S}^{\perp}$.*

## B.2 The Space $\mathsf{L}_2(\Omega, \mathcal{F}, P_\theta)$

Given a fixed parameter $\theta \in \Theta \subset \mathbb{R}^d$, consider the set $\mathcal{L}_2(\Omega, \mathcal{F}, P_\theta)$ of all real-valued random variables $\boldsymbol{x}$, defined over the probability space $(\Omega, \mathcal{F}, P_\theta)$, with zero mean and

$$\mathbb{E}[\boldsymbol{x}^2; \theta] < \infty.$$

Observe that such a set is a vector space over $\mathbb{R}$ with the usual addition of random variables and scalar multiplication. The zero vector is taken as the measurable function which is identically zero over $\Omega$. Furthermore, by Minkowski inequality, it holds that

$$\sqrt{\mathbb{E}[(\boldsymbol{x}+\boldsymbol{y})^2]} \le \sqrt{\mathbb{E}[\boldsymbol{x}^2]} + \sqrt{\mathbb{E}[\boldsymbol{y}^2]}, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{L}_2(\Omega, \mathcal{F}, P_\theta)$$

which implies that $\mathcal{L}_2(\Omega, \mathcal{F}, P_\theta)$ is closed under addition. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{L}_2(\Omega, \mathcal{F}, P_\theta)$ define

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \mathbb{E}[\boldsymbol{x}\boldsymbol{y}; \theta]. \tag{B.1}$$

It is easy to show that this definition satisfies the properties of inner product on the set $\mathcal{L}_2(\Omega, \mathcal{F}, P_\theta)$, except that

$$\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0 \text{ does not imply that } \boldsymbol{x} = 0,$$

but only that

$$P_\theta(\boldsymbol{x} = 0) = 1.$$

If we work instead with the vector space of classes of $P_\theta$-equivalent functions, in the set $\mathcal{L}_2(\Omega, \mathcal{F}, P_\theta)$, which is defined by the equivalence relation

$$\boldsymbol{x}, \boldsymbol{y} \text{ are equivalent if } P_\theta(\boldsymbol{x} = \boldsymbol{y}) = 1,$$

(B.1) becomes an inner product. Let us denote such a set of equivalent classes (or the set of representatives) by $\mathsf{L}_2(\Omega, \mathcal{F}, P_\theta)$.

**Theorem B.2.1** (Hilbert space of random variables)**.** *The inner product space $\mathsf{L}_2(\Omega, \mathcal{F}, P_\theta)$ is complete, and therefore is a Hilbert space.*

For brevity, we will drop the argument of $\mathsf{L}_2(\Omega, \mathcal{F}, P_\theta)$ and refer to the space by the symbol $\mathsf{L}_2$.

**Definition B.2.2** (The projection mapping)**.** *If $\mathcal{S}$ is a closed subspace of $\mathsf{L}_2$, and $\mathrm{id}_{\mathsf{L}_2}$ is the identity map on $\mathsf{L}_2$, the projection $\mathcal{P}_\mathcal{S}$ of $\mathsf{L}_2$ onto $\mathcal{S}$ is defined by*

$$\mathcal{P}_\mathcal{S}\boldsymbol{x} := \hat{\boldsymbol{x}}, \quad \textit{for any } \boldsymbol{x} \in \mathsf{L}_2,$$

*in which $\hat{\boldsymbol{x}}$ is the unique element such that*

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_{\mathsf{L}_2} = \inf_{y \in \mathcal{S}} \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathsf{L}_2}.$$

*The existence and uniqueness of the projection mapping is given by the projection theorem. We also have the complement projection map $(\mathrm{id}_{\mathsf{L}_2} - \mathcal{P}_\mathcal{S})$ mapping $\mathsf{L}_2$ onto $\mathcal{S}^\perp$.*

The projection mapping can be shown to satisfy the following properties

i. $\mathcal{P}_{\mathcal{S}}(\alpha \boldsymbol{x} + \beta \boldsymbol{y}) = \alpha \mathcal{P}_{\mathcal{S}} \boldsymbol{x} + \beta \mathcal{P}_{\mathcal{S}} \boldsymbol{y}, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathsf{L}_2$ and $\alpha, \beta \in \mathbb{R}$.

ii. $\|\boldsymbol{x}\|_{\mathsf{L}_2}^2 = \|\mathcal{P}_{\mathcal{S}} \boldsymbol{x}\|_{\mathsf{L}_2}^2 + \|(id_{\mathsf{L}_2} - \mathcal{P}_{\mathcal{S}})\boldsymbol{x}\|_{\mathsf{L}_2}^2$, for any $\boldsymbol{x} \in \mathsf{L}_2$.

iii. For any given closed subspace $\mathcal{S} \subset \mathsf{L}_2$, every $\boldsymbol{x} \in \mathsf{L}_2$ has a unique representation as a sum of a vector in $\mathcal{S}$ and a vector in $\mathcal{S}^{\perp}$, i.e., $\boldsymbol{x} = \mathcal{P}_{\mathcal{S}} \boldsymbol{x} + (id_{\mathsf{L}_2} - \mathcal{P}_{\mathcal{S}})\boldsymbol{x}$.

iv. $\mathcal{P}_{\mathcal{S}} \boldsymbol{x}_n \to \mathcal{P}_{\mathcal{S}} \boldsymbol{x}$ if $\|\boldsymbol{x}_n - \boldsymbol{x}\|_{\mathsf{L}_2} \to 0$, for any $\{x_n\} \subset \mathsf{L}_2$, and $\boldsymbol{x} \in \mathsf{L}_2$.

v. $\boldsymbol{x} \in \mathcal{S} \iff \mathcal{P}_{\mathcal{S}} \boldsymbol{x} = \boldsymbol{x}$, and $x \in \mathcal{S}^{\perp} \iff \mathcal{P}_{\mathcal{S}} = 0$.

vi. For any $\mathcal{S}_1, \mathcal{S}_2 \subset \mathsf{L}_2$, it holds that $\mathcal{S}_1 \subset \mathcal{S}_2 \iff \mathcal{P}_{\mathcal{S}_1} \mathcal{P}_{\mathcal{S}_2} \boldsymbol{x} = \mathcal{P}_{\mathcal{S}_1} \boldsymbol{x} \quad \forall \boldsymbol{x} \in \mathsf{L}_2$.

**Definition B.2.3** (The prediction equation)**.** *Let $\mathcal{S} \subset \mathsf{L}_2$ be a closed subspace. For any $\boldsymbol{x} \in \mathsf{L}_2$, the equation*

$$\langle \boldsymbol{x} - \hat{\boldsymbol{x}}, \boldsymbol{y} \rangle = 0 \quad \forall \boldsymbol{y} \in \mathcal{S}$$

*defining $\hat{\boldsymbol{x}}$ are known as the predictions equation. Here, $\hat{\boldsymbol{x}}$ is the unique vector defined by the projection theorem. The prediction equation is therefore seen as a restatement of condition (ii) in Theorem B.1.9.*

The above development can be generalized to real vector valued random variables.

## B.3 The Space $\mathsf{L}_2^n(\Omega, \mathcal{F}, P_\theta)$

Consider the set $\mathcal{L}_2^n(\Omega, \mathcal{F}, P_\theta)$ of random vectors $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^\top$, in which $\boldsymbol{x}_i \in \mathsf{L}_2$ for $i = 1, 2, \ldots, n$, for some finite $n \in \mathbb{N}$. Such a set forms a vector space over $\mathbb{R}$ with the usual notion of random vector addition and multiplication by reals. We can introduce the function

$$\langle \boldsymbol{X}, \boldsymbol{Z} \rangle := \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{Z}; \theta] \quad \text{for any } \boldsymbol{X}, \boldsymbol{Z} \in \mathcal{L}_2^n.$$

It is easy to show that such a function satisfies the definition of an inner product on the set of classes of $P_\theta$-equivalent random vectors in $\mathcal{L}_2^n(\Omega, \mathcal{F}, P_\theta)$, and therefore we get an inner product space which we denote $\mathsf{L}_2^n(\Omega, \mathcal{F}, P_\theta)$. This inner product space can be shown to be complete, and therefore it is a Hilbert space. This means that all the statements developed in Section B.2 hold for the space $\mathsf{L}_2^n(\Omega, \mathcal{F}, P_\theta)$, including the projection theorem. For brevity, we will drop the argument of $\mathsf{L}_2^n(\Omega, \mathcal{F}, P_\theta)$ and refer to the space by the symbol $\mathsf{L}_2^n$.

## B.4 Linear Minimum Mean-Square Error Prediction

### B.4.1 Projection in $\mathsf{L}_2$

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ be vectors in $\mathsf{L}_2$. These vectors generate a finite-dimensional subspace

$$D_N := \mathbf{sp}\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\} \subset \mathsf{L}_2.$$

Given an arbitrary vector $\boldsymbol{x} \in \mathsf{L}_2$, we seek a vector $\hat{\boldsymbol{x}} \in D_N$ that best approximates $\boldsymbol{x}$ in the sense of minimizing $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_{\mathsf{L}_2}$. Since the subspace $D_N$ is finite dimensional, it is closed in $\mathsf{L}_2$, and the projection theorem guarantees the existence and uniqueness of the solution. Due to the finite dimensionality of $D_N$, the prediction equations can be easily used to characterize the solution. Since $\hat{\boldsymbol{x}} \in D_N$, we have

$$\hat{\boldsymbol{x}} = \alpha_1 \boldsymbol{y}_1 + \cdots + \alpha_n \boldsymbol{y}_n, \quad \alpha_1, \ldots, \alpha_N \in \mathbb{R}.$$

Therefore, the problem reduces to finding the $N$ scalars $\alpha_i$, $i = 1, \ldots, N$. The prediction equations imply that

$$\begin{aligned}
\alpha_1 \langle \boldsymbol{y}_1, \boldsymbol{y}_1 \rangle + \alpha_2 \langle \boldsymbol{y}_2, \boldsymbol{y}_1 \rangle + \ldots \alpha_N \langle \boldsymbol{y}_n, \boldsymbol{y}_1 \rangle &= \langle \boldsymbol{x}, \boldsymbol{y}_1 \rangle, \\
\alpha_1 \langle \boldsymbol{y}_1, \boldsymbol{y}_2 \rangle + \alpha_2 \langle \boldsymbol{y}_2, \boldsymbol{y}_2 \rangle + \ldots \alpha_N \langle \boldsymbol{y}_n, \boldsymbol{y}_2 \rangle &= \langle \boldsymbol{x}, \boldsymbol{y}_2 \rangle, \\
&\vdots \\
\alpha_1 \langle \boldsymbol{y}_1, \boldsymbol{y}_N \rangle + \alpha_2 \langle \boldsymbol{y}_2, \boldsymbol{y}_N \rangle + \ldots \alpha_N \langle \boldsymbol{y}_n, \boldsymbol{y}_N \rangle &= \langle \boldsymbol{x}, \boldsymbol{y}_N \rangle.
\end{aligned}$$

Define the vectors

$$\begin{aligned}
\alpha &= [\alpha_1, \ldots \alpha_N]^\intercal, \\
\beta &= [\langle \boldsymbol{x}, \boldsymbol{y}_1 \rangle, \langle \boldsymbol{x}, \boldsymbol{y}_2 \rangle, \ldots \langle \boldsymbol{x}, \boldsymbol{y}_N \rangle]^\intercal, \\
\boldsymbol{Y} &= [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]^\intercal,
\end{aligned}$$

and the matrix

$$\Sigma = \begin{bmatrix} \langle \boldsymbol{y}_1, \boldsymbol{y}_1 \rangle & \langle \boldsymbol{y}_2, \boldsymbol{y}_1 \rangle & \cdots & \langle \boldsymbol{y}_N, \boldsymbol{y}_1 \rangle \\ \langle \boldsymbol{y}_1, \boldsymbol{y}_2 \rangle & \langle \boldsymbol{y}_2, \boldsymbol{y}_2 \rangle & \cdots & \langle \boldsymbol{y}_N, \boldsymbol{y}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{y}_1, \boldsymbol{y}_N \rangle & \langle \boldsymbol{y}_2, \boldsymbol{y}_N \rangle & \cdots & \langle \boldsymbol{y}_N, \boldsymbol{y}_N \rangle \end{bmatrix}.$$

Then, finding $\hat{\boldsymbol{x}}$ is equivalent to solving, for $\alpha$, the system of linear equations

$$\Sigma \alpha = \beta.$$

The projection is given by

$$\hat{\boldsymbol{x}} = \boldsymbol{Y}^\intercal \begin{bmatrix} \langle \boldsymbol{y}_1, \boldsymbol{y}_1 \rangle & \langle \boldsymbol{y}_2, \boldsymbol{y}_1 \rangle & \cdots & \langle \boldsymbol{y}_N, \boldsymbol{y}_1 \rangle \\ \langle \boldsymbol{y}_1, \boldsymbol{y}_2 \rangle & \langle \boldsymbol{y}_2, \boldsymbol{y}_2 \rangle & \cdots & \langle \boldsymbol{y}_N, \boldsymbol{y}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{y}_1, \boldsymbol{y}_N \rangle & \langle \boldsymbol{y}_2, \boldsymbol{y}_N \rangle & \cdots & \langle \boldsymbol{y}_N, \boldsymbol{y}_N \rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle \boldsymbol{x}, \boldsymbol{y}_1 \rangle \\ \langle \boldsymbol{x}, \boldsymbol{y}_2 \rangle \\ \vdots \\ \langle \boldsymbol{x}, \boldsymbol{y}_N \rangle \end{bmatrix}. \tag{B.2}$$

### B.4.2 Projection in $\mathsf{L}_2^n$

Let a vector $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]^\intercal$ with entries in $\mathsf{L}_2$ be given and consider the subspace

$$S = \mathbf{sp}\{\ \{\boldsymbol{y}_i^{(j)}\} : i = 1, \ldots, N, \text{ and } j = 1, \ldots, n\},$$

in which $\boldsymbol{y}_i^{(j)} \in \mathsf{L}_2^n$ is the vector valued random variable with the $j^{\text{th}}$ entry equal to $\boldsymbol{y}_i$ and all other entries are equal to the zero vector of $\mathsf{L}_2$. The subspace $S$ is a finite dimensional subspace of $\mathsf{L}_2^n$. It can be equivalently represented by multiplying the vector $\boldsymbol{Y}$ with an arbitrary $n \times N$ matrix, that is to say

$$S = \{L\boldsymbol{Y} : L \text{ is an arbitrary } n \times N \text{ matrix of real numbers}\} \subset \mathsf{L}_2^n.$$

We are interested in finding the projection of an arbitrary element

$$\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]^\top \in \mathsf{L}_2^n$$

onto $S$. We assume that $\boldsymbol{X} \notin S$. The projection theorem guarantees the existence and uniqueness of a projection $\hat{\boldsymbol{X}} = L_* \boldsymbol{Y}$ that satisfies the orthogonality condition $(\boldsymbol{X} - \hat{\boldsymbol{X}}) \perp S$, or

$$\langle \boldsymbol{X} - L_* \boldsymbol{Y}, L\boldsymbol{Y} \rangle = 0 \text{ for all matrices } L.$$

It is not difficult to show (by expanding the matrix multiplication and choosing $L$ appropriately) that this is equivalent to

$$\mathbb{E}[(\boldsymbol{X} - L_* \boldsymbol{Y})\boldsymbol{Y}^\top; \theta] = 0$$

where the zero is the $n \times N$ zero matrix and therefore, the projection matrix $L_*$ is given by

$$L_* = \mathbf{cov}(\boldsymbol{X}, \boldsymbol{Y})\Sigma^{-1},$$

in which $\Sigma = \mathbf{cov}(\boldsymbol{Y}, \boldsymbol{Y})$. The projection is given by

$$\hat{\boldsymbol{X}}^\top = \boldsymbol{Y}^\top \begin{bmatrix} \langle \boldsymbol{y}_1, \boldsymbol{y}_1 \rangle & \langle \boldsymbol{y}_2, \boldsymbol{y}_1 \rangle & \cdots & \langle \boldsymbol{y}_N, \boldsymbol{y}_1 \rangle \\ \langle \boldsymbol{y}_1, \boldsymbol{y}_2 \rangle & \langle \boldsymbol{y}_2, \boldsymbol{y}_2 \rangle & \cdots & \langle \boldsymbol{y}_N, \boldsymbol{y}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{y}_1, \boldsymbol{y}_N \rangle & \langle \boldsymbol{y}_2, \boldsymbol{y}_n \rangle & \cdots & \langle \boldsymbol{y}_N, \boldsymbol{y}_N \rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle \boldsymbol{x}_1, \boldsymbol{y}_1 \rangle & \langle \boldsymbol{x}_2, \boldsymbol{y}_1 \rangle & \cdots & \langle \boldsymbol{x}_n, \boldsymbol{y}_1 \rangle \\ \langle \boldsymbol{x}_1, \boldsymbol{y}_2 \rangle & \langle \boldsymbol{x}_2, \boldsymbol{y}_2 \rangle & \cdots & \langle \boldsymbol{x}_n, \boldsymbol{y}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{x}_1, \boldsymbol{y}_N \rangle & \langle \boldsymbol{x}_2, \boldsymbol{y}_N \rangle & \cdots & \langle \boldsymbol{x}_n, \boldsymbol{y}_N \rangle \end{bmatrix}.$$

This comes in agreement with the obtained results in [67] where a different space is used with matrix valued inner products. Comparing with [67, Theorem 3.2.1 in page 81], we see that the projection $\hat{\boldsymbol{X}}$ does not only minimize $\|\boldsymbol{X} - \boldsymbol{Z}\|_{\mathsf{L}_2^n}$ over all $\boldsymbol{Z} \in S$, but also minimizes the error covariance matrix

$$\mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{Z})(\boldsymbol{X} - \boldsymbol{Z})^\top; \theta\right]$$

over all $\boldsymbol{Z} \in S$.

Comparing this result with the results of the scalar case in (B.2), we see that the projection of a vector valued random variable $\boldsymbol{X} \in \mathsf{L}_2^n$ is given by the projections (in $\mathsf{L}_2$) of its individual entries $\boldsymbol{x}_i$ for $i = 1, \dots, n$ onto $\mathbf{D}_N$, then stacking them together in one column vector.

## B.5 Summary

In this appendix, we reviewed the definition and some properties of Hilbert spaces of random variables. The main result that is relevant to the material of this thesis is the projection theorem. It guarantees the existence and uniqueness of a linear minimum mean-square estimator of a random variable given a set of correlated but arbitrary random variables with finite second order moments. Due to its linearity, such an estimator relies only on the first and second order moments of the involved random variables.

Appendix C

# The Multivariate Gaussian Distribution

In this appendix, we gather some results of multivariate Gaussian random variables. The material is standard and can be found in [2], [67], or [139] for example. The importance of multivariate Gaussian distributions stems from its mathematical properties. For example, the Gaussian family of distributions is closed under affine transformations and their marginalization integrals can be computed analytically. A Gaussian distribution is also completely determined by the the mean vector and the covariance matrix. Moreover, it appears as a fundamental limiting distribution in central limit theorems.

## C.1 Multivariate Gaussian Random Variables

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]^\top$ be a vector valued random variable. We say that $\boldsymbol{X}$ is a Gaussian random variable and write

$$\boldsymbol{X} \sim \mathcal{N}(\mu(\theta), \Sigma(\theta)),$$

if its PDF is

$$p(X;\theta) = \frac{1}{(2\pi)^{n/2}[\det \Sigma(\theta)]^{1/2}} \exp\left(-\frac{1}{2}(X - \mu(\theta))^\top \Sigma^{-1}(\theta)(X - \mu(\theta))\right)$$

for a vector $\mu(\theta) \in \mathbb{R}^n$ and an $n \times n$ matrix $\Sigma(\theta) \succ 0$. We say that $\boldsymbol{X}$ is a standard Gaussian random variable if

$$\boldsymbol{X} \sim \mathcal{N}(0, I).$$

Let $\boldsymbol{X} \sim \mathcal{N}(\mu(\theta), \Sigma(\theta))$ take values in $\mathbb{R}^n$. Then it holds that

$$\mathbb{E}[\boldsymbol{X};\theta] = \mu(\theta), \quad \text{and} \quad \mathbf{cov}(\boldsymbol{X}, \boldsymbol{X};\theta) = \Sigma(\theta).$$

Define

$$\boldsymbol{Y} = A\boldsymbol{X} + b$$

in which $b \in \mathbb{R}^p$ and $A$ is a matrix of dimension $p \times n$ with full row rank. Then it is easy to show that

$$\boldsymbol{Y} \sim \mathcal{N}(A\mu(\theta) + b, A\Sigma(\theta)A^\top).$$

Because $\boldsymbol{X}$ and $\boldsymbol{Y}$ are related linearly, it also holds that they are jointly Gaussian. This means that the vector

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X}^\top & \boldsymbol{Y}^\top \end{bmatrix}^\top$$

is a Gaussian random variable.

## C.2 Conditional Distribution of Multivariate Gaussian Random Variables

The conditional distribution of jointly Gaussian random vectors is also a Gaussian distribution. It is completely characterized by the conditional mean and the conditional covariance matrix.

**Theorem C.2.1.** *Consider a partitioned Gaussian random vector*

$$Z = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{bmatrix} \right).$$

*Then it holds that*

$$\begin{aligned} \boldsymbol{X} &\sim \mathcal{N}(\mu_1, \Sigma_1), \\ \boldsymbol{Y} &\sim \mathcal{N}(\mu_2, \Sigma_2), \\ \boldsymbol{X}|Y &\sim \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_2^{-1}(Y - \mu_2)\ ,\ \Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21}), \ \textit{and} \\ \boldsymbol{Y}|X &\sim \mathcal{N}(\mu_2 + \Sigma_{21}\Sigma_1^{-1}(X - \mu_1)\ ,\ \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12}). \end{aligned}$$

Observe that the conditional expectation

$$\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y};\theta] = \mu_1 + \Sigma_{12}\Sigma_2^{-1}(\boldsymbol{Y} - \mu_2)$$

coincides with the projection $\hat{\boldsymbol{X}}$ of $(\boldsymbol{X} - \mu_1) \in \mathsf{L}_2^n$ onto the subspace spanned by the entries of $(\boldsymbol{Y} - \mu_2)$ in $\mathsf{L}_2^n$, see Appendix B. Therefore, the linear minimum mean-squares error estimator of $\boldsymbol{X}$ given $\boldsymbol{Y}$ coincides with the unconstrained minimum mean-square error estimator.

# Bibliography

[1] T. Amemiya. *Advanced Econometrics.* Harvard University Press, 1985.

[2] B. D. O. Anderson and J. B. Moore. *Optimal filtering.* Prentice Hall, 1979.

[3] R. Ash and C. Doléans-Dade. *Probability and Measure Theory.* Academic Press, 2000.

[4] S. Asmussen and P. Glynn. *Stochastic Simulation: Algorithms and Analysis.* Springer, 2007.

[5] K. J. Åström and T. Bohlin. Numerical identification of linear dynamic systems from normal operating records. In *Theory of Self-Adaptive Control Systems*, pages 96–111. 1966.

[6] K. Åström and B. Wittenmark. *Computer-Controlled Systems: Theory and Design, Third Edition.* Dover Publications, 2013.

[7] K. Åström. *Introduction to Stochastic Control Theory.* Academic Press, 1970.

[8] R. R. Bahadur. On Fisher's bound for asymptotic variances. *The Annals of Mathematical Statistics*, 35(4):1545–1552, 1964.

[9] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141, 1988.

[10] G. Bastin and D. Dochain. Chapter 1- Dynamical models of bioreactors. In G. Bastin, , and D. Dochain, editors, *On-line Estimation and Adaptive Control of Bioreactors*, pages 1 – 82. Elsevier, 1990.

[11] J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):179–195, 1975.

[12] S. A. Billings. Identification of nonlinear systems- a survey. *IEE Proceedings D - Control Theory and Applications*, 1980.

[13] S. A. Billings. *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains.* Wiley, 2013.

[14] G. Birpoutsoukis and J. Schoukens. Nonparametric Volterra kernel estimation using regularization. In *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, pages 222–227, 2015.

[15] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[16] R. A. Boyles. On the convergence of the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(1):47–50, 1983.

[17] G. Burgers, P. J. van Leeuwen, and G. Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126(6):1719–1724, 1998.

[18] J. Bussgang. Cross-correlation functions of amplitude-distorted Gaussian signals. Technical report, Research Laboratory of Electronics, Massachusetts Institute of Technology, 1952.

[19] P. Caines. *Linear stochastic systems.* J. Wiley, 1988.

[20] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models.* Springer, 2006.

[21] O. Cappé. Online sequential Monte Carlo EM algorithm. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 37–40, 2009.

[22] B. P. Carlin, N. G. Polson, and D. S. Stoffer. A Monte Carlo approach to non-normal and nonlinear state-space modeling. *Journal of the American Statistical Association*, 87(418):493–500, 1992.

[23] K. Chung. *A Course in Probability Theory.* Academic Press, 2001.

[24] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 2012.

[25] H. Cox. On the estimation of state variables and parameters for noisy dynamic systems. *IEEE Transactions on Automatic Control*, 9(1):5–12, 1964.

[26] H. Cramér. A contribution to the theory of statistical estimation. *Scandinavian Actuarial Journal*, 1946(1):85–94, 1946.

[27] H. Cramér. On some classes of nonstationary stochastic processes. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Contributions to Probability Theory*, pages 57–78, Berkeley, Calif., 1961. University of California Press.

[28] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[30] L. Devroye. *Non-Uniform Random Variate Generation.* Springer, 1986.

[31] J. Doob. *Stochastic processes.* Wiley, 1953.

[32] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

[33] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 2000.

[34] A. S. Eddington. The evaluation of the cosmical number. *Mathematical Proceedings of the Cambridge Philosophical Society*, 40(1):37–56, 1944.

[35] B. Efron. Why isn't everyone a Bayesian? *The American Statistician*, 40(1):1–5, 1986.

[36] M. Enqvist. *Linear Models of Nonlinear Systems.* Linköping Studies in Science and Technology. Dissertations No. 985, Institutionen för systemteknik, Linköping University, 2005.

[37] M. Evans and T. Swartz. *Approximating Integrals via Monte Carlo and Deterministic Methods.* Oxford University Press, 2000.

[38] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994.

[39] G. Evensen. The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control Systems*, 29(3):83–104, 2009.

[40] R. A. Fisher. On an absolute criterion for fitting frequency curves. *Messenger of Mathmatics*, 41:155–160, 1912.

[41] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 222(594-604):309–368, 1922.

[42] R. Fletcher. *Practical Methods of Optimization.* Wiley, 2000.

[43] G. Fort and E. Moulines. Convergence of the Monte Carlo Expectation Maximization for curved exponential families. *The Annals of Statistics*, 31(4):1220–1259, 2003.

[44] J. Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339, 1989.

[45] S. Gillijns, O. B. Mendoza, J. Chandrasekar, B. L. R. D. Moor, D. S. Bernstein, and A. Ridley. What is the ensemble Kalman filter and how well does it work? In *2006 American Control Conference*, 2006.

[46] G. Giordano and J. Sjöberg. Consistency aspects of Wiener-Hammerstein model identification in presence of process noise. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 3042–3047. 2016.

[47] C. Giraud. *Introduction to high-dimensional statistics.* CRC Press, 2014.

[48] F. Giri and E. Bai. *Block-oriented Nonlinear System Identification.* Springer, 2010.

[49] P. Glasserman and D. D. Yao. Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908, 1992.

[50] P. W. Glynn and D. L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392, 1989.

[51] G. Golub and C. Van Loan. *Matrix Computations.* Johns Hopkins University Press, 1996.

[52] G. C. Goodwin and R. L. Payne. *Dynamic System Identification: Experiment Design and Data Analysis.* Academic Press, 1977.

[53] G. Goodwin and K. Sin. *Adaptive Filtering Prediction and Control.* Dover Publications, 2014.

[54] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing*, 140(2):107–113, 1993.

[55] C. Gouriéroux, A. Monfort, and A. Trognon. Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3):681–700, 1984.

[56] L. Grippo and M. Sciandrone. Nonmonotone globalization techniques for the Barzilai-Borwein gradient method. *Computational Optimization and Applications*, 23(2):143–169, 2002.

[57] R. Haber and H. Unbehauen. Structure identification of nonlinear dynamic systems - a survey on input/output approaches. *Automatica*, 26(4):651 – 677, 1990.

[58] A. Hagenblad, L. Ljung, and A. Wills. Maximum likelihood identification of Wiener models. *Automatica*, 44(11):2697 – 2705, 2008.

[59] J. Hammersley and D. Handscomb. *Monte Carlo Methods.* Methuen, 1965.

[60] E. Hannan and M. Deistler. *The statistical theory of linear systems.* Wiley, 1988.

[61] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1994.

[62] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 2012.

[63] A. J. Isaksson, J. Sjöberg, D. Törnqvist, L. Ljung, and M. Kok. Using horizon estimation and nonlinear optimization for grey-box identification. *Journal of Process Control*, 30:69 – 79, 2015.

[64] A. Jazwinski. *Stochastic Processes and Filtering Theory.* Dover Publications, 2007.

[65] R. Johansson. *System Modeling and Identification.* Prentice Hall, 1993.

[66] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte. A new approach for filtering nonlinear systems. In *American Control Conference, Proceedings of the 1995*, Vol. 3, pages 1628–1632, 1995.

[67] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation.* Prentice Hall, 2000.

[68] T. Kailath. *Linear Systems.* Prentice Hall, 1980.

[69] R. Kalman, P. Falb, and M. Arbib. *Topics in mathematical system theory.* McGraw-Hill, 1969.

[70] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[71] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin. On particle methods for parameter estimation in state-space models. *Statist. Sci.*, 30(3):328–351, 2015.

[72] G. Kitagawa. Non-Gaussian state-space modeling of nonstationary time series: Rejoinder. *Journal of the American Statistical Association*, 82(400):1060–1063, 1987.

[73] A. Klenke. *Probability Theory: A Comprehensive Course.* Springer, 2013.

[74] D. Knuth. *The Art of Computer Programming: Volume 1: Fundamental Algorithms.* Pearson Education, 1997.

[75] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on Optimization*, 9(1):112–147, 1998.

[76] K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):425–437, 1995.

[77] K. Lange. A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica*, 5(1):1–18, 1995.

[78] P. S. Laplace. Memoir on the probability of the causes of events. *Mémoires de Mathématique et de Physique, Tome Sixième*, 1774. (English translation by S. M. Stigler 1986. *Statist. Sci.*, 1(19):364-378).

[79] L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related results. *University of California Publications in Statistics*, 1:277–330, 1953.

[80] L. Le Cam. Maximum likelihood: An introduction. *International Statistical Review / Revue Internationale de Statistique*, 58(2):153–171, 1990.

[81] F. Le Gland, V. Monbet, and V.-D. Tran. Large sample asymptotics for the ensemble Kalman filter. Research Report RR-7014, INRIA, 2009.

[82] E. L. Lehmann and G. Casella. *Theory of Point Estimation.* Springer, 2011.

[83] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses.* Springer, 2006.

[84] E. L. Lehmann. *Elements of Large-Sample Theory.* Springer, 1999.

[85] F. Lindsten, M. I. Jordan, and T. B. Schön. Particle Gibbs with ancestor sampling. *Journal of Machine Learning Research*, 15(1):2145–2184, 2014.

[86] F. Lindsten. An efficient stochastic approximation EM algorithm using conditional particle filters. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6274–6278, 2013.

[87] F. Lindsten. *Particle filters and Markov chains for learning of dynamical systems.* Linköping Studies in Science and Technology. Dissertations No. 1530, Linköping University, 2013.

[88] J. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, 2008.

[89] L. Ljung and P. E. Caines. Asymptotic normality of prediction error estimators for approximate system models. *Stochastics*, 3(1-4):29–46, 1980.

[90] L. Ljung. Convergence analysis of parametric identification methods. *IEEE Transactions on Automatic Control*, 23(5):770–783, 1978.

[91] L. Ljung. A non-probabilistic framework for signal spectra. In *1985 24th IEEE Conference on Decision and Control*, pages 1056–1060, 1985.

[92] L. Ljung. *System Identification: Theory for the User.* Prentice Hall, 2nd edition, 1999.

[93] L. Ljung. Estimating linear time-invariant models of nonlinear time-varying systems. *European Journal of Control*, 7(2):203 – 219, 2001.

[94] D. G. Luenberger and Y. Ye. *Linear and Nonlinear Programming.* Springer, 2015.

[95] D. G. Luenberger. *Introduction to Dynamic Systems: Theory, Models, and Applications.* Wiley, 1979.

[96] S. N. Maceachern, M. Clyde, and J. S. Liu. Sequential importance sampling for nonparametric Bayes models: The next generation. *Canadian Journal of Statistics*, 27(2):251–267, 1999.

[97] A. Marconato, J. Sjöberg, J. Suykens, and J. Schoukens. Separate initialization of dynamics and nonlinearities in nonlinear state-space models. In *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pages 2104–2108, 2012.

[98] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley, 2007.

[99] X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[100] M. Milanese. *Bounding approaches to system identification.* Plenum Press, 1996.

[101] G. Mzyk. *Combined Parametric-Nonparametric Identification of Block-Oriented Systems.* Springer, 2013.

[102] R. C. Neath. *On Convergence Properties of the Monte Carlo EM Algorithm*, Vol. 10 of *Collections*, pages 43–62. Institute of Mathematical Statistics, 2013.

[103] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[104] O. Nelles. *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models.* Springer, 2001.

[105] S. F. Nielsen. The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489, 2000.

[106] B. Ninness, A. Wills, and T. Schön. Estimation of general nonlinear state-space systems. In *49th IEEE Conference on Decision and Control, Atlanta, Georgia, USA*, pages 1–6, 2010.

[107] B. Ninness. Some system identification challenges and approaches. *IFAC Proceedings Volumes*, 42(10):1 – 20, 2009.

[108] J. Nocedal and S. Wright. *Numerical Optimization.* Springer, 2006.

[109] J. Olsson, O. Cappé, R. Douc, and E. Moulines. Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1):155–179, 2008.

[110] J. Paduart, L. Lauwers, J. Swevers, K. Smolders, J. Schoukens, and R. Pintelon. Identification of nonlinear systems using polynomial nonlinear state space models. *Automatica*, 46(4):647 – 656, 2010.

[111] R. Pintelon and J. Schoukens. *System Identification: A Frequency Domain Approach.* Wiley, 2nd edition, 2012.

[112] D. N. Politis. Computer-intensive methods in statistical analysis. *IEEE Signal Processing Magazine*, 15(1):39–55, 1998.

[113] C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Cal. Math. Soc.*, 37(3):81–91, 1945.

[114] J. Rawlings and D. Mayne. *Model Predictive Control: Theory and Design.* Nob Hill Pub., 2009.

[115] F. Riesz and B. Nagy. *Functional Analysis.* Dover Publications, 2012.

[116] B. Ripley. *Stochastic Simulation.* Wiley, 2009.

[117] C. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, 2013.

[118] C. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation.* Springer, 2007.

[119] M. Roth, C. Fritsche, G. Hendeby, and F. Gustafsson. The ensemble Kalman filter and its relations to other nonlinear filters. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1236–1240, 2015.

[120] I. Rozanov. *Stationary random processes.* Holden-Day, 1967.

[121] R. Y. Rubinstein. *Simulation and the Monte Carlo Method.* John Wiley & Sons, 1981.

[122] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, 1987.

[123] M. Schervish. *Theory of Statistics.* Springer, 1996.

[124] M. Schetzen. *The Volterra and Wiener theories of nonlinear systems.* Wiley, 1980.

[125] T. B. Schön, A. Wills, and B. Ninness. System identification of nonlinear state-space models. *Automatica*, 47(1):39 – 49, 2011.

[126] T. B. Schön, F. Lindsten, J. Dahlin, J. Wågberg, C. A. Naesseth, A. Svensson, and L. Dai. Sequential Monte Carlo methods for system identification. *IFAC-PapersOnLine*, 48(28):775 – 786, 2015.

[127] T. B. Schön. *Estimation of nonlinear dynamic systems: Theory and applications.* Linköping Studies in Science and Technology. Dissertations No. 998, Institutionen för systemteknik, Linköping University, 2006.

[128] M. Schoukens and K. Tiels. Identification of nonlinear block-oriented systems starting from linear approximations: A survey. *CoRR*, abs/1607.01217, 2016.

[129] J. Schoukens, A. Marconato, R. Pintelon, Y. Rolain, M. Schoukens, K. Tiels, L. Vanbeylen, G. Vandersteen, and A. V. Mulders. System identification in a real world. In *2014 IEEE 13th International Workshop on Advanced Motion Control (AMC)*, 2014.

[130] J. Schoukens, M. Vaes, and R. Pintelon. Linear system identification in a nonlinear setting: Nonparametric analysis of the nonlinear distortions and their impact on the best linear approximation. *IEEE Control Systems*, 36(3):38–69, 2016.

[131] M. Schoukens, R. Pintelon, T. Dobrowiecki, and J. Schoukens. Extending the best linear approximation framework to the process noise case. A Poster presented at the 25th workshop of the European Research Network System Identification (ERNSI), Cison di Valmarino, Italy., September 2016.

[132] M. Schoukens. *Identification of parallel block-oriented models starting from the best linear approximation.* PhD thesis, Vrije Universiteit Brussel, 2015.

[133] J. Sjöberg and J. Schoukens. Initializing wiener-hammerstein models based on partitioning of the best linear approximation. *Automatica*, 48(2):353 – 359, 2012.

[134] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31(12):1691 – 1724, 1995.

[135] J. Sjöberg, L. Lauwers, and J. Schoukens. Identification of wiener-hammerstein models: Two algorithms based on the best split of a linear model applied to the sysid'09 benchmark problem. *Control Engineering Practice*, 20(11):1119 – 1125, 2012.

[136] J. Sjöberg. On estimation of nonlinear black-box models: how to obtain a good initialization. In *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*, pages 72–81, 1997.

[137] S. Skogestad and I. Postlethwaite. *Multivariable Feedback Control: Analysis and Design.* Wiley, 2005.

[138] T. Söderström and P. Stoica. *System Identification.* Prentice Hall, 1989.

[139] T. Söderström. *Discrete-time Stochastic Systems: Estimation and Control.* Springer, 2002.

[140] S. M. Stigler. The epic story of maximum likelihood. *Statistical Science*, 22(4):598–620, 2007.

[141] L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[142] M. Verhaegen and V. Verdult. *Filtering and System Identification: A Least Squares Approach.* Cambridge University Press, 2012.

[143] B. Wahlberg. *On the Identification and Approximation of Linear Systems.* Linköping Studies in Science and Technology. Dissertations No. 163, Linköping University, 1987.

[144] G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

[145] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series.* Wiley, 1949.

[146] A. Wills, T. B. Schön, L. Ljung, and B. Ninness. Blind identification of Wiener models. *IFAC Proceedings Volumes*, 44(1):5597 – 5602, 2011.

[147] A. Wills, T. B. Schön, L. Ljung, and B. Ninness. Identification of Hammerstein-Wiener models. *Automatica*, 2013.

[148] H. Wold. *A Study in the Analysis of Stationary Time Series.* Almqvist & Wiksells boktrycheri-a.-b., 1938.

[149] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[150] A. Yeredor. The joint MAP-ML criterion and its relation to ML and to extended least-squares. *IEEE Transactions on Signal Processing*, 48(12):3484–3492, 2000.

[151] N. Young. *An Introduction to Hilbert Space.* Cambridge University Press, 1988.

[152] L. Zadeh and E. Polak. *System Theory.* McGraw-Hill, 1969.

[153] K. Zhou, J. Doyle, and K. Glover. *Robust and Optimal Control.* Prentice Hall, 1996.

[154] Y. Zhu. *Multivariable System Identification For Process Control.* Elsevier Science, 2001.