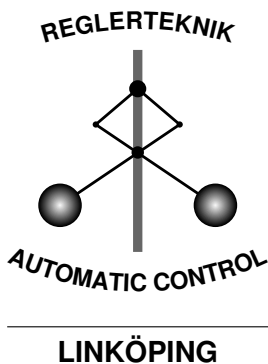


Linköping Studies in Science and Technology  
Thesis No. 1199

# Fundamental Estimation and Detection Limits in Linear Non-Gaussian Systems

Gustaf Hendeby



Division of Automatic Control  
Department of Electrical Engineering  
Linköpings universitet, SE-581 83 Linköping, Sweden  
<http://www.control.isy.liu.se>  
[hendeby@isy.liu.se](mailto:hendeby@isy.liu.se)

Linköping 2005

This is a Swedish Licentiate's Thesis.  
Swedish postgraduate education leads to a Doctor's degree and/or a Licentiate's degree. A Doctor's Degree comprises 160 credits (4 years of full-time studies). A Licentiate's degree comprises 80 credits, of which at least 40 credits constitute a Licentiate's thesis.

**Fundamental  
Estimation and Detection Limits  
in Linear Non-Gaussian Systems**

© 2005 Gustaf Hendeby

*Department of Electrical Engineering  
Linköpings universitet  
SE-581 83 Linköping  
Sweden*

ISBN 91-85457-40-X

ISSN 0280-7971

LiU-TEK-LIC-2005:54

Printed by LiU-Tryck, Linköping, Sweden 2005

*Unintentionally left blank!*



## Abstract

Many methods used for estimation and detection consider only the mean and variance of the involved noise instead of the full noise descriptions. One reason for this is that the mathematics is often considerably simplified this way. However, the implications of the simplifications are seldom studied, and this thesis shows that if no approximations are made performance is gained. Furthermore, the gain is quantified in terms of the useful information in the noise distributions involved. The useful information is given by the *intrinsic accuracy*, and a method to compute the intrinsic accuracy for a given distribution, using Monte Carlo methods, is outlined.

A lower bound for the covariance of the estimation error for any unbiased estimator is given by the *Cramér-Rao lower bound* (CRLB). At the same time, the *Kalman filter* is the *best linear unbiased estimator* (BLUE) for linear systems. It is in this thesis shown that the CRLB and the BLUE performance are given by the same expression, which is parameterized in the intrinsic accuracy of the noise. How the performance depends on the noise is then used to indicate when nonlinear filters, *e.g.*, a particle filter, should be used instead of a Kalman filter. The CRLB results are shown, in simulations, to be a useful indication of when to use more powerful estimation methods. The simulations also show that other techniques should be used as a complement to the CRLB analysis to get conclusive performance results.

For fault detection, the statistics of the asymptotic *generalized likelihood ratio* (GLR) test provides an upper bound on the obtainable detection performance. The performance is in this thesis shown to depend on the intrinsic accuracy of the involved noise. The asymptotic GLR performance can then be calculated for a test using the actual noise and for a test using the approximative Gaussian noise. Based on the difference in performance, it is possible to draw conclusions about the quality of the Gaussian approximation. Simulations show that when the difference in performance is large, an exact noise representation improves the detection. Simulations also show that it is difficult to predict the exact influence on the detection performance caused by substituting the system noise with Gaussian noise approximations.



# Sammanfattning

Många metoder som används i estimerings- och detekteringssammanhang tar endast hänsyn till medelvärde och varians hos ingående brus istället för att använda en fullständig brusbeskrivning. En av anledningarna till detta är att den förenklade brusmodellen underlättar många beräkningar. Dock studeras sällan de effekter förenklarna leder till. Denna avhandling visar att om inga förenklingar görs kan prestandan förbättras och det visas också hur förbättringen kan relateras till den intressanta informationen i det involverade bruset. Den intressanta informationen är den inneboende noggrannheten (eng. *intrinsic accuracy*) och ett sätt för att bestämma den inneboende noggrannheten hos en given fördelning med hjälp av Monte-Carlo-metoder presenteras.

Ett mått på hur bra en estimator utan bias kan göras ges av *Cramér-Raos undre gräns* (CRLB). Samtidigt är det känt att *kalmanfiltret* är den bästa lineära biasfria estimatorn (BLUE) för lineära system. Det visas här att CRLB och BLUE-prestanda ges av samma matematiska uttryck där den inneboende noggrannheten ingår som en parameter. Kunskap om hur informationen påverkar prestandan kan sedan användas för att indikera när ett olineärt filter, t.ex. ett partikelfilter, bör användas istället för ett kalmanfilter. Med hjälp av simuleringar visas att CRLB är ett användbart mått för att indikera när mer avancerade metoder kan vara lönsamma. Simuleringarna visar dock också att CRLB-analysen bör kompletteras med andra tekniker för att det ska vara möjligt att dra definitiva slutsatser.

I fallet feldetektion ger de asymptotiska egenskaperna hos den generaliserade sannolikhetskvoten (eng. *generalized likelihood ratio*, GLR) en övre gräns för hur bra detektorer som kan konstrueras. Det visas här hur den övre gränsen beror på den inneboende noggrannheten hos det aktuella bruset. Genom att beräkna asymptotisk GLR-testprestanda för det sanna bruset och för en gaussisk brusapproximation går det att dra slutsatser om huruvida den gaussiska approximationen är tillräckligt bra för att kunna användas. I simuleringar visas att det är lönsamt att använda sig av en exakt brusbeskrivning om skillnaden i prestanda är stor mellan de båda fallen. Simuleringarna indikerar också att det kan vara svårt att förutsäga den exakta effekten av en gaussisk brusapproximation.



# Acknowledgments

Here I am — finally! I never thought this day would ever come. Once here, I want to thank everyone who helped me get here today, starting with my parents and sisters who have always been there for me, encouraging me in my interest in science and to pursue my dreams. I also want to thank Sara Rassner who, even though she is a biologist, has been a great friend and support throughout many late night chats.

I'm immensely thankful to everyone that over the years have had the guts to believe in me. Lately, Prof. Lennart Ljung, who let me join his research group and who has ever since been there to make sure that I've found my place, and my supervisor Prof. Fredrik Gustafsson, who has taken me under his wings. Fredrik has since the very first day I sat my foot here been an endless source of ideas, inspiration, and interesting discussions. I hope I haven't disappointed you too much so far. Not to forget, Ulla Salaneck — you are the best! Ulla, without your help with everything practical the group would not survive. Keep up the good work, we appreciate it!

I would also like to thank everyone who has proofread this thesis, especially Dr. Rickard Karlsson who has offered constant support during the writing of this thesis and always seems to be able to come up with invaluable comments and suggestions. Other helpful proofreaders have been Jeroen Hol, Dr. Mikael Norrlöf, Lic. Thomas Schön, Johan Sjöberg, and David Törnqvist, who have all read various parts of the thesis and helped me to improve it. Without your help guys, this thesis would not be what it is today.

I'm in debt to the entire Automatic Control group in Linköping because without the endless coffee breaks I'd gone crazy years ago. A special thank you to Lic. Erik Geijer Lundin, you have been a constant source of laughter during odd hours at the office, especially the last couple of weeks. Thank you everyone who have tried to limit my ambitions for my pet projects, especially Mikael, Lic. Martin Enqvist, and Rickard — I know it hasn't been easy, but you've really made a difference. I also want to thank the people who joined the group at the same time that I did, David, Johan, Daniel Axehill, and later Henrik Tidefelt, for getting me through the first year and keeping me going ever since with interesting conversations and encouragement.

Last, but not least important, this work has been sponsored by VINNOVA's Center of Excellence ISIS (Information Systems for Industrial Control and Supervision) at Linköpings universitet, Linköping, Sweden. Thank you!

Linköping, October 2005  
*Gustaf Hendebý*



---

# Contents

- 1 Introduction** **1**
  - 1.1 Research Motivation . . . . . 2
  - 1.2 Problem Formulation . . . . . 3
  - 1.3 Contributions . . . . . 3
  - 1.4 Thesis Outline . . . . . 4
  
- 2 Statistics** **5**
  - 2.1 Stochastic Variables . . . . . 5
    - 2.1.1 Distribution Definition . . . . . 5
    - 2.1.2 Information in Distributions . . . . . 7
    - 2.1.3 Kullback-Leibler Measures . . . . . 10
  - 2.2 Monte Carlo Integration . . . . . 12
    - 2.2.1 Intrinsic Accuracy using Monte Carlo Integration . . . . . 13
    - 2.2.2 Kullback Divergence using Monte Carlo Integration . . . . . 13
  - 2.3 Studied Distributions . . . . . 14
    - 2.3.1 Gaussian Distribution . . . . . 14
    - 2.3.2 Gaussian Mixture Distribution . . . . . 15
  - 2.4 Transformed Distributions . . . . . 19
    - 2.4.1 Monte Carlo Transformation . . . . . 21
    - 2.4.2 Gauss Approximation Formula . . . . . 22
    - 2.4.3 Unscented Transform . . . . . 22
  
- 3 Models** **27**
  - 3.1 State-Space Model . . . . . 27
    - 3.1.1 General State-Space Model . . . . . 28
    - 3.1.2 Linear State-Space Model . . . . . 28
    - 3.1.3 Model with Fault Terms . . . . . 29

---

3.1.4	Batched Linear State-Space Model . . . . .	30
3.2	Hidden Markov Model . . . . .	31
<b>4</b>	<b>Filtering Methods</b>	<b>33</b>
4.1	Parameter Estimation . . . . .	34
4.2	Particle Filter . . . . .	35
4.2.1	Approximative Probability Density Function . . . . .	35
4.2.2	Resampling . . . . .	38
4.3	Kalman Filter . . . . .	39
4.4	Kalman Filters for Nonlinear Models . . . . .	40
4.4.1	Linearized Kalman Filter . . . . .	41
4.4.2	Extended Kalman Filter . . . . .	41
4.4.3	Iterated Extended Kalman Filter . . . . .	43
4.5	Unscented Kalman Filter . . . . .	43
4.6	Filter Banks . . . . .	46
4.6.1	Complete Filter Bank . . . . .	47
4.6.2	Filter Bank with Pruning . . . . .	48
4.7	Comparison of Nonlinear Filtering Methods . . . . .	51
4.7.1	Problem Description . . . . .	51
4.7.2	Studied Methods . . . . .	52
4.7.3	Monte Carlo Simulations . . . . .	56
<b>5</b>	<b>Cramér-Rao Lower Bound</b>	<b>59</b>
5.1	Parametric Cramér-Rao Lower Bound . . . . .	59
5.2	Posterior Cramér-Rao Lower Bound . . . . .	61
5.3	Cramér-Rao Lower Bounds for Linear Systems . . . . .	62
5.4	Applications of the Theory . . . . .	67
5.4.1	Constant Velocity Model . . . . .	67
5.4.2	DC Motor . . . . .	70
5.4.3	Observations . . . . .	73
<b>6</b>	<b>Change Detection</b>	<b>75</b>
6.1	Hypothesis Testing . . . . .	75
6.2	Test Statistics . . . . .	78
6.2.1	Likelihood Ratio Test . . . . .	78
6.2.2	Generalized Likelihood Ratio Test . . . . .	80
6.3	Most Powerful Detector . . . . .	80
6.4	Asymptotic Generalized Likelihood Ratio Test . . . . .	81
6.4.1	Wald Test . . . . .	82
6.4.2	Detection Performance . . . . .	83
6.5	Uniformly Most Powerful Test for Linear Systems . . . . .	85
6.5.1	Linear System Residuals . . . . .	85
6.5.2	Prior initial state knowledge . . . . .	87
6.5.3	Parity Space . . . . .	88
6.6	Applications of the Theory . . . . .	88
6.6.1	Constant Velocity Model . . . . .	89

6.6.2	DC Motor . . . . .	92
6.6.3	Observations . . . . .	94
<b>7</b>	<b>Concluding Remarks</b>	<b>97</b>
7.1	Conclusions . . . . .	97
7.2	Further Work . . . . .	98
<b>A</b>	<b>Notational Conventions</b>	<b>101</b>
	<b>Bibliography</b>	<b>105</b>



# 1

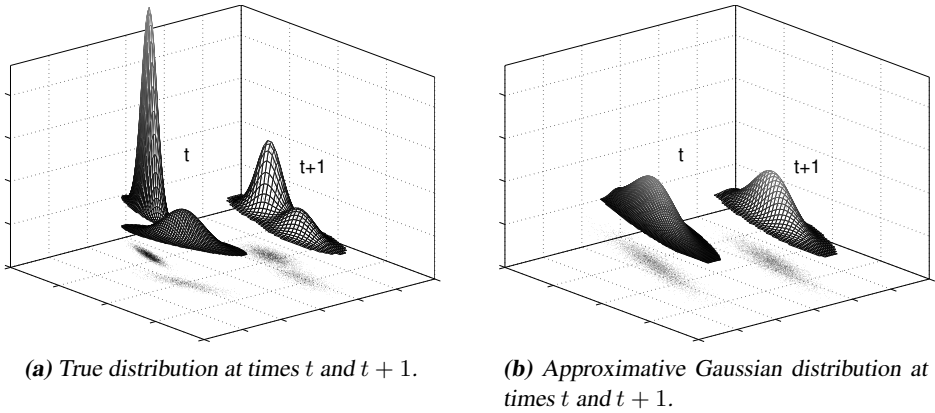
---

## Introduction

**E**LECTRONICS WITH seemingly intelligent behavior have become an important part of everyday in life the last couple of years, and the trend is towards even more advanced technology. One example is cars; most new cars are equipped with *adaptive cruise control* (ACC), *anti-lock braking systems* (ABS), *traction control systems* (TCS), and sometimes systems to locate the car and to give the driver directions. In the future, systems to avoid or at least lessen the effect of collisions will be included into cars to enhance safety. Another example where new technology is introduced is navigation at sea, where recent development is towards using radar measurements to navigate instead of the global positioning system (GPS) which is vulnerable in many respects.

All the examples given above have in common that they rely heavily on information about their surroundings. Unfortunately, such information is usually expensive to obtain since it requires sensors. For situations where the price is an important factor, this is a real problem and producers try their best to use as few and as inexpensive sensors as possible. How the information from the sensor is processed is therefore of utter importance. With proper usage of the measured information there is much to be gained. In part, this thesis deals with determining how much information is available from sensor measurements, and the motivation is the wish utilize sensor data as much as possible. From the theory developed in this thesis it is possible to tell when it is worthwhile to spend time on advanced signal processing and when all sensor information available is extracted and put to use. For estimation this is done in terms of the Cramér-Rao lower bound, and for change detection in terms of the asymptotic properties of the generalized likelihood ratio test.

A result of having information intensive products is that they tend to be sensitive to hardware failures and other faults. With less expensive and simpler sensors, probably of lower quality, the risk of malfunction increases, and if the same sensor is used for several different purposes this could potentially be a major problem. Suppose the sensor is used for both an anti-collision system and a braking system, if the sensor malfunctions the consequences could easily be fatal. Therefore, sensors and other parts in these systems



(a) True distribution at times  $t$  and  $t + 1$ .

(b) Approximative Gaussian distribution at times  $t$  and  $t + 1$ .

**Figure 1.1:** Distribution of the states in a system developing over time, left to right, for the true distribution and a Gaussian approximation. Samples from the distributions are included for reference.

need to be carefully monitored, so that when a fault is detected, countermeasures could be taken to minimize the damage and to put the system into a fail-safe mode. This thesis deals with determining the potential to detect faults in a system for a given sensor configuration. The results can then be used as an upper bound on what is possible to do, and to serve as a guide for design and implementation.

This chapter introduces the problem studied in this thesis with a short motivation, followed by a stricter problem formulation. After that the contributions to this thesis are listed and the content of the thesis is outlined.

## 1.1 Research Motivation

In many situations nonlinear systems are linearized and all noise assumed Gaussian for practical reasons. Basically, linear Gaussian systems are easy to handle. However, introducing approximations into the system description, both in terms of linearizations and Gaussian approximations, comes at a price in terms of performance. The performance loss is often ignored, but this is not always a good idea.

To illustrate how the information may be lost, *e.g.*, quantified in terms of intrinsic accuracy, when Gaussian approximations are used, consider the two probability density functions in Figure 1.1(a). They can for instance be said to represent the position of a car or ship at two different time instances. Especially the first distribution is distinctly bimodal, which shows as a clear clustering in the samples from the distribution. Figure 1.1(b) illustrates the Gaussian approximations of the very same distributions. The approximations have different characteristics compared to the true distribution, and as a consequence information is lost. For instance, the region separating the modes in the initial distribution, is for the Gaussian distribution considered likely, and quite a few samples are found there in the approximation. This behavior shows up as a much higher intrinsic

accuracy for the true distribution than for the Gaussian approximation. It is from this example not difficult to see that given the Gaussian approximation of this distribution, it is not possible to derive as much information about the state as from the correct distribution.

Well-known methods to utilize available information in non-Gaussian distributions and nonlinear systems exist, e.g., the particle filter. However, the price for using all available information may be high. An increase in need for computational resources is often mentioned, but it is also likely that the process of developing customized methods is more difficult than using standardized methods, e.g., an extended Kalman filter or a Gaussian GLR test. For these reasons, and others, linearization and Gaussian approximations are often used. To make better use of data, it would be valuable to have methods to beforehand determine when standard methods are adequate and to have rules of thumb to help design nonstandard filters and detectors.

## 1.2 Problem Formulation

Traditionally, estimation and change detection problems are often treated as being linear and affected by Gaussian noise. When this is not the case, linearization and noise approximations are used to create approximative linear and Gaussian systems. However, the effects on performance introduced this way is often ignored, even though methods exist to handle nonlinear non-Gaussian systems. The reasons for this are many, e.g., the computational complexity increases with more general methods and the design choices are different from the classical ones.

It would be ideal to have a framework to, from a complete system description, give guidelines for when classical approximations are appropriate and when other methods should be used. With such guidelines, the design effort can be put to use with methods appropriate for the specific problem at hand. Furthermore, the framework could also help in other design choices, e.g., give rules of thumb to tune parameters and choose between different parameterizations and system descriptions. This thesis handles linear systems with non-Gaussian noise and tries to determine when nonlinear estimation and detection methods should be used, based on the information available from the noise in the system.

## 1.3 Contributions

The material in this thesis is based on [33–35], but also on material not yet published elsewhere.

The first contribution is a preliminary study on the importance of non-Gaussian noise for estimation:

Gustaf Hendeby and Fredrik Gustafsson. On performance measures for approximative parameter estimation. In *Proceedings of Reglermöte 2004*, Chalmers, Gothenburgh, Sweden, May 2004.

Most of the material in this paper appears in modified form in Chapter 2, where information in distributions are discussed.

The analysis of the Cramér-Rao lower bound of linear systems with non-Gaussian noise, mainly presented in Chapter 5, is an extension to the work initiated in [8] and appeared in:

Gustaf Hendeby and Fredrik Gustafsson. Fundamental filtering limitations in linear non-Gaussian systems. In *Proceedings of 16th Triennial IFAC World Congress*, Prague, Czech Republic, July 2005.

The main contribution is to derive explicit expressions for optimal estimation performance, in terms of the Cramér-Rao lower bound, compared to the quality of the best linear unbiased estimate.

The material on detection performance, presented in Chapter 6, derives a uniform residual formulation for linear systems. Based on this, it is shown how the information content in the involved noise affects the fault detectability in terms of the test statistics of the asymptotic *generalized likelihood ratio* (GLR) test. Chapter 6 is based on the paper:

Gustaf Hendeby and Fredrik Gustafsson. Fundamental fault detection limitations in linear non-Gaussian systems. In *Proceedings of 44th IEEE Conference on Decision and Control and European Control Conference*, Sevilla, Spain, December 2005. To appear.

## 1.4 Thesis Outline

The thesis is separated in four main parts. The first part gives the fundamentals needed to read the rest of the thesis; the statistics of noise and information related measures are introduced in Chapter 2, and in Chapter 3 the different classes of models used in the thesis are introduced to the reader.

Estimation and optimal estimation performance are treated in the second part. Chapter 4 outlines several estimation algorithms for linear and nonlinear systems, and a brief study of characteristics of the different methods for a bearings-only problem is presented. General bounds for the performance of state estimation are given in Chapter 5, as well as specific results and simulations for linear systems affected by non-Gaussian noise.

The third part shifts the focus from state estimation to change detection, and Chapter 6 presents change detection and fundamental performance limitations. The chapters about estimation and detection can be read independently. As for state estimation, general results are given and then applied to linear systems. The results are exemplified in simulations.

Chapter 7 concludes the thesis. Abbreviations and other notation is introduced in the text throughout the thesis as they are needed, but can also be found in Appendix A.

# 2

---

## Statistics

**N**ATURE IS AT THE same time very predictable and very unpredictable. Many phenomena are predictable on the large scale, but unpredictable when it comes to details. For example, a dropped paper will fall to the floor, however exactly when and where it will land is almost impossible to tell beforehand. Much of the paper's behavior is predictable; it is affected by gravity and the surrounding air according to the laws of physics, but a gust of wind caused by a sudden movement in the room may be impossible to predict. The wind gust can be treated as a random or *stochastic* input and the laws of physics is a model for the behavior. This chapter first covers the basics of stochastic variables; fundamentals, properties, and ways to handle transformations. The theory provides the mathematics to express uncertain elements.

### 2.1 Stochastic Variables

*Stochastic variables* (SV) are used in models of systems to describe unknown input. The stochastic input is often called *noise*. This section introduces noise from a statistical point of view and serves as a foundation for the work in the rest of the thesis.

#### 2.1.1 Distribution Definition

A stochastic variable is a variable without a specific value, that has different values with certain probabilities. To describe this behavior a *distribution function* or *cumulative distribution function* (CDF) is used. The CDF for the stochastic variable  $X$  is given by

$$P_X(x) = \Pr(X < x), \quad (2.1)$$

where  $\Pr(\mathcal{A})$  denotes the probability that the statement  $\mathcal{A}$  is true. Hence,  $P_X(x)$  denotes the probability that  $X$  is less than  $x$ . It follows from the definition of probability, that

$$\lim_{x \rightarrow -\infty} P_X(x) = 0, \quad \lim_{x \rightarrow +\infty} P_X(x) = 1,$$

and that  $P_X(x)$  is nondecreasing in  $x$ . Any function that fulfills these three conditions is a CDF and defines a statistical distribution.

Another distribution description, more frequently used in this thesis, is the *probability density function* (PDF),

$$p_X(x) = \frac{dP_X(x)}{dx}, \quad (2.2)$$

which describes the likelihood of certain  $X$  values, *i.e.*,

$$\Pr(x \in \mathcal{S}) = \int_{\mathcal{S}} p_X(x) dx.$$

Discrete stochastic variables are defined in a similar way.

When the behavior of  $X$  is conditioned on another variable  $Y$  this is indicated by

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}, \quad (2.3)$$

where  $p_{X|Y}(x|y)$  is the conditional PDF of  $x$  given the information  $y$ .

The CDF and the PDF both contain a complete description of the underlying stochastic variable. However, the following properties are often studied to get a better understanding of the behavior of the variable:

**Expected value** — the average value of several samples from the distribution,

$$\mu_X = \mathbb{E}(X) = \int xp_X(x) dx. \quad (2.4a)$$

When needed, an index will be used on  $\mathbb{E}$  to clarify which distribution to use in the integral, *e.g.*,  $\mathbb{E}_x$  or  $\mathbb{E}_{p_X(x)}$  depending on the circumstances.

In some situations it is necessary to compute the expected value of a function of a stochastic variable. It is done using the integral

$$\mathbb{E}(f(X)) = \int f(x)p_X(x) dx.$$

**Variance** or *covariance matrix* for multidimensional distributions — indicates how close to the mean a sample can be expected to be,

$$\Sigma_X = \text{cov}(X) = \mathbb{E}((X - \mu_X)(X - \mu_X)^T), \quad (2.4b)$$

or  $\Sigma_X = \text{var}(x)$  for the scalar case. The same index system will be used as for expected values.

**Skewness** has no trivial extension to multidimensional distributions — indicates if samples are expected to have a symmetric behavior around the mean,

$$\gamma_{1,X} = \mathbb{E} \left( X^3 \right) \Sigma_X^{-\frac{3}{2}}. \quad (2.4c)$$

**Kurtosis** has no trivial extensions to multidimensional distributions — gives information about how heavy tails the distribution has,

$$\gamma_{2,X} = \mathbb{E} \left( X^4 \right) \Sigma_X^{-2} - 3. \quad (2.4d)$$

Traditionally, and still in Fisher based statistics, uppercase variables are used to denote stochastic variables, and their lowercase counterparts are used for realizations of them. However, this thesis will from here on use lower case variables in both cases, as in the Bayesian framework unless there is a risk of confusion.

### 2.1.2 Information in Distributions

Samples from a stochastic variable can be more or less informative about underlying parameters of the distribution. A measure of just how much information can be extracted about a parameter is given by the *Fisher information* (FI). As a special case of the Fisher information, the *intrinsic accuracy* (IA) measures the information available from samples from the distribution to determine the expected value. Fisher information and intrinsic accuracy are described in this section, and the relative measure *relative accuracy* (RA) is defined.

#### Fisher Information

The concept of *Fisher information* was first introduced by Fisher in [21] and then further elaborated on in [22] as he was trying to quantify how well parameters of a stochastic variable could be estimated from samples from the distribution.

**Definition 2.1.** The *Fisher information* (FI) is defined [47], under the mild regularity condition that the PDF  $p(x|\theta)$  must have twice continuous partial derivatives, as

$$\mathcal{I}_x(\theta) := -\mathbb{E}_x \left( \Delta_\theta^\theta \log p(x|\theta) \right) = \mathbb{E}_x \left( \left( \nabla_\theta \log p(x|\theta) \right) \left( \nabla_\theta \log p(x|\theta) \right)^T \right) \quad (2.5)$$

evaluated for the true parameter value  $\theta = \theta^0$ , with  $\nabla_x$  and  $\Delta_x^x$  defined in Appendix A to be the *gradient* and the *Hessian*, respectively.

The Fisher information for multidimensional parameters is sometimes called the *Fisher information matrix* to more clearly indicate that it is a matrix, however in this thesis Fisher information is used to denote both.

The inverse of the Fisher information represents a lower bound on the variance of any unbiased estimate,  $\hat{\theta}(x)$ , of  $\theta$ , where  $\theta$  is a parameter to be estimated from the measurements  $x$ . More precisely [47, 58],

$$\text{cov} \left( \hat{\theta}(x) \right) \succeq \mathcal{I}_x^{-1}(\theta),$$

where  $A \succeq B$  denotes that the matrix  $A - B$  is positive semidefinite, i.e.,  $x^T(A - B)x \geq 0$  for all  $x$  of suitable dimension. This lower bound on the variance of an estimated parameter, introduced by Rao [70], is often referred to as the *Cramér-Rao lower bound* (CRLB) [47],

$$P_{\theta}^{\text{CRLB}} = \mathcal{I}_x^{-1}(\theta).$$

The CRLB will, when extended to handle dynamic systems, play an important role in this thesis. For now, just note the connection between the Fisher information and the CRLB.

### Intrinsic Accuracy

The Fisher information with respect to the expected value,  $\mu$ , of a stochastic variable is extra important, therefore, introduce the short hand notation  $\mathcal{I}_x := \mathcal{I}_x(\mu)$ . This quantity is in [15, 48, 49] referred to as the *intrinsic accuracy* (IA) of the PDF for  $x$ . This name is inspired by the terminology used in Fisher [22]. Intrinsic accuracy can be expressed as

$$\begin{aligned} \mathcal{I}_x &= -\mathbb{E}_x\left(\Delta_{\mu}^{\mu} \log p_x(x|\mu)\Big|_{\mu=\mu^0}\right) = -\mathbb{E}_x\left(\Delta_{\mu}^{\mu} \log p_x(x - \mu|0)\Big|_{\mu=\mu^0}\right) \\ &= -\mathbb{E}_x\left(\Delta_x^x \log p_x(x - \mu|0)\Big|_{\mu=\mu^0}\right) = -\mathbb{E}_x\left(\Delta_x^x \log p_x(x|\mu)\Big|_{\mu=\mu^0}\right), \end{aligned} \quad (2.6)$$

where the two middle steps follows because  $\mu$  is the mean of the distribution. Analogously it can be shown using the second form of the expression for Fisher information that

$$\mathcal{I}_x = \mathbb{E}_x\left(\left(\nabla_x \log p_x(x|\mu)\right)\left(\nabla_x \log p_x(x|\mu)\right)^T \Big|_{\mu=\mu^0}\right). \quad (2.7)$$

When estimating the mean of a stochastic variable it is always possible to achieve an unbiased estimator with the same covariance as the stochastic variable, and in some cases it is possible to get an even better estimator. In fact, for non-Gaussian distributions the lower bound is always better as stated in the following theorem.

#### Theorem 2.1

*For the intrinsic accuracy and covariance of the stochastic variable  $x$  the following holds*

$$\text{cov}(x) \succeq \mathcal{I}_x^{-1},$$

*with equality if and only if  $x$  is Gaussian.*

**Proof:** See [76]. □

In this respect the Gaussian distribution is a worst case distribution. Of all distributions with the same covariance the Gaussian distribution is the distribution with the least information about its mean. All other distributions have larger intrinsic accuracy.

### Relative Accuracy

The relation between intrinsic accuracy and covariance is interesting in many situations. In most cases only the relative difference between the two matters, therefore introduce *relative accuracy* (RA).

**Definition 2.2.** If a scalar  $\Psi_x$  exists such that  $\text{cov}(x) = \Psi_x \mathcal{I}_x^{-1}$ , denote  $\Psi_x$  the *relative accuracy* for the distribution.

It follows from Theorem 2.1 that when the relative accuracy is defined,  $\Psi_x \geq 1$ , with equality if and only if  $x$  is Gaussian. The relative accuracy is thus a measure of how much useful information there is in the distribution, compared to a Gaussian distribution with the same covariance.

### Accuracy Properties

The intrinsic accuracy for independent variables are separable. This is intuitive and can be used to simplify calculations considerably.

#### Lemma 2.1

For a vector  $\mathbb{X}$  of independently distributed stochastic variables  $\mathbb{X} = (x_1^T, \dots, x_n^T)^T$  each with covariance  $\text{cov}(x_i) = \Sigma_{x_i}$  and intrinsic accuracy  $\mathcal{I}_{x_i}$ , for  $i = 1, \dots, n$ ,

$$\text{cov}(\mathbb{X}) = \text{diag}(\Sigma_{x_1}, \dots, \Sigma_{x_n}) \quad \text{and} \quad \mathcal{I}_{\mathbb{X}} = \text{diag}(\mathcal{I}_{x_1}, \dots, \mathcal{I}_{x_n}).$$

If  $\Sigma_{x_i} = \Sigma_x$  and  $\mathcal{I}_{x_i} = \mathcal{I}_x$  then the covariance and the intrinsic accuracy are more compactly expressed

$$\text{cov}(\mathbb{X}) = I \otimes \Sigma_x \quad \text{and} \quad \mathcal{I}_{\mathbb{X}} = I \otimes \mathcal{I}_x,$$

where  $\otimes$  denotes the Kronecker product. Furthermore, if  $\Sigma_x \mathcal{I}_x = \Psi_x \cdot I$ , with  $\Psi_x \geq 1$  a scalar, then

$$\text{cov}(\mathbb{X}) = \Psi_x \mathcal{I}_{\mathbb{X}}^{-1} = \Psi_x I \otimes \mathcal{I}_x^{-1}.$$

**Proof:** For  $\mathbb{X} = (x_1^T, x_2^T, \dots, x_n^T)^T$ , with  $x_i$  independently distributed, it follows immediately that

$$\text{cov}(\mathbb{X}) = \text{diag}(\text{cov}(x_1), \dots, \text{cov}(x_n)) = \text{diag}(\Sigma_{x_1}, \dots, \Sigma_{x_n}).$$

Expand the expression:

$$\mathcal{I}_{\mathbb{X}} = -\text{E} \left( \Delta_{\mathbb{X}}^{\mathbb{X}} \log p(\mathbb{X}) \right) = -\text{E} \left( \Delta_{\mathbb{X}}^{\mathbb{X}} \log \prod_{i=1}^n p(x_i) \right) = \sum_{i=1}^n -\text{E} \left( \Delta_{\mathbb{X}}^{\mathbb{X}} \log p(x_i) \right).$$

The partial derivatives of this expression becomes, for  $k = l = i$ ,

$$-\text{E} \left( \nabla_{x_k} \nabla_{x_l} \log p(x_i) \right) = -\text{E} \left( \Delta_{x_i}^{x_i} \log p(x_i) \right) = \mathcal{I}_{x_i},$$

and for  $k \neq l$  the partial derivatives vanish,  $-\text{E} \left( \nabla_{x_k} \nabla_{x_l} \log p(x_i) \right) = 0$ . Combining these results using matrix notation yields

$$\mathcal{I}_{\mathbb{X}} = \text{diag}(\mathcal{I}_{x_1}, \dots, \mathcal{I}_{x_n}).$$

The compact notation for  $\Sigma_{x_i} = \Sigma_x$  and  $\mathcal{I}_{x_i} = \mathcal{I}_x$  follows as in the proof of Theorem 2.1.  $\square$

Intrinsic accuracy of linear combinations of stochastic variables can be calculated from the intrinsic accuracy of the components.

### Theorem 2.2

For the linear combination of stochastic variables  $\mathbb{Z} = B\mathbb{X}$ , where  $\mathbb{X} = (x_1^T, \dots, x_n^T)^T$  is a stochastic variable with independent components with covariance  $\text{cov}(x_i) = \Sigma_{x_i}$  and intrinsic accuracy  $\mathcal{I}_{x_i}$ ,

$$\begin{aligned}\text{cov}(\mathbb{Z}) &= B \text{diag}(\Sigma_{x_1}, \dots, \Sigma_{x_n}) B^T, \\ \mathcal{I}_{\mathbb{Z}}^{-1} &= B \text{diag}(\mathcal{I}_{x_1}^{-1}, \dots, \mathcal{I}_{x_n}^{-1}) B^T.\end{aligned}$$

**Proof:** Combine the result found as Theorem 4.3 in [8] with Lemma 2.1.  $\square$

If the combined variables are *identically and independently distributed* (IID) the expressions can be further simplified using the last part of Lemma 2.1.

## 2.1.3 Kullback-Leibler Measures

The *Kullback-Leibler information* [54, 55], also called the *discriminating information*, is quantifies the difference between two distributions. The Kullback-Leibler information is not symmetric in its arguments, and therefore not a measure. If a measure is needed, the *Kullback divergence*, constructed as a symmetric sum of two Kullback-Leibler informations [4, 54], can be used as an alternative.

**Definition 2.3.** The *Kullback-Leibler information* is defined, for the two proper PDFs  $p(\cdot)$  and  $q(\cdot)$ , as

$$\mathcal{I}^{\text{KL}}(p(\cdot), q(\cdot)) = \int p(x) \log \frac{p(x)}{q(x)} dx, \quad (2.8a)$$

when  $p(x) \neq 0 \Leftrightarrow q(x) \neq 0$  and otherwise as  $\mathcal{I}^{\text{KL}}(p(\cdot), q(\cdot)) = +\infty$ . The *Kullback divergence* is from this defined to be

$$\mathcal{J}^{\text{K}}(p(\cdot), q(\cdot)) = \mathcal{I}^{\text{KL}}(p(\cdot), q(\cdot)) + \mathcal{I}^{\text{KL}}(q(\cdot), p(\cdot)), \quad (2.8b)$$

under the same restrictions on  $p(\cdot)$  and  $q(\cdot)$ .

The Kullback-Leibler information is closely related to other statistical measures, e.g., *Shannon's information* and *Akaike's information criterion* [4]. A connection to the Fisher information can also be found [54].

Both the Kullback-Leibler information and the Kullback divergence are additive for independent stochastic variables as shown in Lemma 2.2.

### Lemma 2.2

For a vector  $\mathbb{X}$  of independent stochastic variables,  $\mathbb{X} = (x_1^T, \dots, x_n^T)^T$  distributed with PDF  $p(\mathbb{X}) = \prod_{i=1}^n p_i(x_i)$  or  $q(\mathbb{X}) = \prod_{i=1}^n q_i(x_i)$  the *Kullback-Leibler information* and *Kullback divergence* are additive, i.e.,

$$\mathcal{I}^{\text{KL}}(p(\cdot), q(\cdot)) = \sum_{i=1}^n \mathcal{I}^{\text{KL}}(p_i(\cdot), q_i(\cdot))$$

and

$$\mathcal{J}^{\mathbf{K}}(p(\cdot), q(\cdot)) = \sum_{i=1}^n \mathcal{J}^{\mathbf{K}}(p_i(\cdot), q_i(\cdot)).$$

**Proof:** If  $\mathbb{X}$ ,  $p(\cdot)$ , and  $q(\cdot)$  satisfy the conditions in the lemma, then

$$\begin{aligned} \mathcal{I}^{\text{KL}}(p(\cdot), q(\cdot)) &= \int p(\mathbb{X}) \log \frac{p(\mathbb{X})}{q(\mathbb{X})} d\mathbb{X} \\ &= \int \prod_{j=1}^n p_j(x_j) \log \frac{\prod_{i=1}^n p_i(x_i)}{\prod_{i=1}^n q_i(x_i)} d\mathbb{X} = \sum_{i=1}^n \int \prod_{j=1}^n p_j(x_j) \log \frac{p_i(x_i)}{q_i(x_i)} d\mathbb{X} \\ &= \sum_{i=1}^n \int p_i(x_i) \log \frac{p_i(x_i)}{q_i(x_i)} dx_i = \sum_{i=1}^n \mathcal{I}^{\text{KL}}(p_i(\cdot), q_i(\cdot)), \end{aligned}$$

where the second last equality is a *marginalization* utilizing that  $p_i(\cdot)$ , for all  $i$ , are proper PDFs and hence integrate to unity.

Now the Kullback divergence result follows immediately,

$$\begin{aligned} \mathcal{J}^{\mathbf{K}}(p(\cdot), q(\cdot)) &= \mathcal{I}^{\text{KL}}(p(\cdot), q(\cdot)) + \mathcal{I}^{\text{KL}}(q(\cdot), p(\cdot)) \\ &= \sum_{i=1}^n \mathcal{I}^{\text{KL}}(p_i(\cdot), q_i(\cdot)) + \sum_{i=1}^n \mathcal{I}^{\text{KL}}(q_i(\cdot), p_i(\cdot)) \\ &= \sum_{i=1}^n \left( \mathcal{I}^{\text{KL}}(p_i(\cdot), q_i(\cdot)) + \mathcal{I}^{\text{KL}}(q_i(\cdot), p_i(\cdot)) \right) = \sum_{i=1}^n \mathcal{J}^{\mathbf{K}}(p_i(\cdot), q_i(\cdot)). \end{aligned}$$

□

## Theoretical Bounds

There exists bounds on both the Kullback-Leibler information and the Kullback divergence in terms of measures that are easier to compute. The bound in Theorem 2.3 is one of the sharpest lower bounds known.

### Theorem 2.3

A lower bound for the Kullback-Leibler information is defined in terms of the variational distance,  $V(\cdot, \cdot)$ , between the two PDFs  $p(\cdot)$  and  $q(\cdot)$ ,

$$V(p(\cdot), q(\cdot)) = \int |p(x) - q(x)| dx,$$

and is

$$\begin{aligned} \mathcal{I}^{\text{KL}}(p(\cdot), q(\cdot)) &\geq \max \left\{ L_1 \left( V(p(\cdot), q(\cdot)) \right), L_2 \left( V(p(\cdot), q(\cdot)) \right) \right\} \\ L_1 \left( V(p(\cdot), q(\cdot)) \right) &= \log \frac{2 + V(p(\cdot), q(\cdot))}{2 - V(p(\cdot), q(\cdot))} - \frac{2V(p(\cdot), q(\cdot))}{2 - V(p(\cdot), q(\cdot))} \\ L_2 \left( V(p(\cdot), q(\cdot)) \right) &= \frac{V^2(p(\cdot), q(\cdot))}{2} + \frac{V^4(p(\cdot), q(\cdot))}{36} + \frac{V^6(p(\cdot), q(\cdot))}{288}, \end{aligned}$$

for  $0 \leq V(p(\cdot), q(\cdot)) \leq 2$ .

**Proof:** See [62]. □

Unfortunately there seems to be few useful upper bounds for the Kullback-Leibler information and the Kullback divergence.

## 2.2 Monte Carlo Integration

*Monte Carlo integration* [73] is a method to use statistical properties to compute integrals that are otherwise hard to handle. Basically, the idea is to reformulate difficult integrals on a form where computing an expected value renders the integral of interest. To illustrate this, consider the integral

$$I := \int f(x) dx = \int g(x)p(x) dx,$$

where  $p(\cdot)$  should be a proper PDF and  $g(x) := f(x)/p(x)$ . The value of the integral,  $I$ , can then be approximated with the sum

$$\hat{I}_N := \sum_{i=1}^N \frac{1}{N} g(x^{(i)}),$$

where  $\{x^{(i)}\}_{i=1}^N$  are  $N$  IID samples, or *particles*, from the distribution given by  $p(\cdot)$ . The approximation utilizes that  $I = \mathbb{E}_{p(x)} g(x)$  and that an expected value can be approximated with a sample mean. Furthermore, it follows from the *law of large numbers* that if  $\text{var}_{p(x)}(g(x)) = \Sigma$ , and  $\Sigma$  is bounded, then

$$\lim_{N \rightarrow +\infty} \sqrt{N}(\hat{I}_N - I) \sim \mathcal{N}(0, \Sigma),$$

i.e.,  $\hat{I}_N \rightarrow I$  as  $N \rightarrow +\infty$  and the quality of the estimate improves with increasing  $N$  [19, 73]. Note, the convergence is in theory independent of the state-space dimension, and Monte Carlo integration should hence suffer little from the *curse of dimensionality* in contrast to the deterministic integration methods. However, this is according to [17, 67] overly optimistic and Monte Carlo integration is claimed to suffer from the curse of dimensionality. This, on the other hand, seems too pessimistic for most applications in practice.

If may be difficult, or even impossible, to draw samples from  $p(\cdot)$ . This is sometimes the case with the *a posteriori* state distributions used later. If this is the problem, choose another proper PDF  $q(\cdot)$  such that  $p(x) > 0$  implies  $q(x) > 0$  for  $x$  in the domain of  $p(\cdot)$ , i.e., the support of  $p(\cdot)$  is included in the support of  $q(\cdot)$ . Using  $q(\cdot)$  in the approximation yields,

$$\hat{I}_N = \sum_{i=1}^N \frac{p(x^{(i)})}{Nq(x^{(i)})} g(x^{(i)}) = \sum_{i=1}^N \omega^{(i)} g(x^{(i)}),$$

with the same limit and principal convergence as before. The distribution given by  $q(\cdot)$  is often called an *importance distribution* and  $\omega^{(i)}$  *importance weights*. Note that even if

$p(\cdot)$ , and thus  $\omega^{(i)}$ , is only known up to a normalizing constant, this is not a problem since

$$\sum_i \omega^{(i)} \rightarrow \int \frac{cp(x)}{q(x)} q(x) dx = \int cp(x) dx = c,$$

and it is hence possible to normalize the distribution and compute the integral anyhow,

$$\hat{I}_N = \frac{\sum_i \omega^{(i)} g(x^{(i)})}{\sum_i \omega^{(i)}}.$$

### 2.2.1 Intrinsic Accuracy using Monte Carlo Integration

It is often difficult to calculate the intrinsic accuracy analytically, however to use Monte Carlo integration is one possible solution. The intrinsic accuracy is defined by

$$\begin{aligned} \mathcal{I}_x &= \mathbb{E}_x \left( (\nabla_x \log p_x(x|\mu)) (\nabla_x \log p_x(x|\mu))^T \Big|_{\mu=\mu^0} \right) \\ &= \int (\nabla_x \log p_x(x|\mu)) (\nabla_x \log p_x(x|\mu))^T p_x(x|\mu) \Big|_{\mu=\mu^0} dx \end{aligned} \quad (2.9)$$

which fits well into the Monte Carlo integration framework with samples from  $p(x|\mu^0)$ . If samples from the distribution  $p(x|\mu^0)$  are available and it is possible to compute

$$\nabla_x \log p_x(x|\mu) = \frac{\nabla_x p_x(x|\mu)}{p(x|\mu)}, \quad (2.10)$$

for given  $\mu$  and  $x$ , the application of the Monte Carlo integration is a straightforward way to compute the intrinsic accuracy. The alternative formulation of the intrinsic accuracy

$$\mathcal{I}_x = -\mathbb{E}_x \left( \Delta_x^x \log p_x(x|\mu) \Big|_{\mu=\mu^0} \right), \quad (2.11)$$

can also be used.

The number of particles that are needed to get an acceptable approximation of the intrinsic accuracy for a distribution must be assessed independently for each distribution. However, before applying Monte Carlo integration it is always a good idea to try to reduce the size of the problem using the relations derived in Section 2.1.2. In most cases reducing the complexity and dimension of the stochastic variable will speed up the computations by limiting the degrees of freedom, the time it takes to draw random numbers, and to compute the probabilities and their derivatives.

### 2.2.2 Kullback Divergence using Monte Carlo Integration

Just as with the intrinsic accuracy it is often hard to calculate the Kullback-Leibler information and the Kullback divergence analytically, but the computations are well suited for Monte Carlo integration. To compute the Kullback-Leibler information,

$$\mathcal{I}^{\text{KL}}(p(\cdot), q(\cdot)) = \int p(x) \log \frac{p(x)}{q(x)} dx, \quad (2.12)$$

samples from the distribution  $p(\cdot)$  are needed and

$$\log \frac{p(x)}{q(x)} = \log p(x) - \log q(x) \quad (2.13)$$

must be computed for the samples. Once again, it is preferable to utilize any rules that lessens the complexity (see Section 2.1.3) of the computations to get results at a lower computational cost. The Kullback-Leibler information can then be used to compute the Kullback divergence if necessary.

## 2.3 Studied Distributions

In this section the *Gaussian distribution* is introduced as one of the most widely used distribution and properties for it are derived. Furthermore, the class of *Gaussian Mixture distributions* is introduced as a complement to the Gaussian distribution.

### 2.3.1 Gaussian Distribution

The most widely used stochastic distribution is the *Gaussian distribution*, or *Normal distribution*, denoted with  $\mathcal{N}(\mu, \Sigma)$ , where  $\mu$  and  $\Sigma \succ 0$  are parameters representing expected value and variance (covariance matrix) of the distribution. The Gaussian distribution is for a scalar stochastic variable defined in terms of its PDF,

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{2\pi\Sigma}} e^{-\frac{(x-\mu)^2}{2\Sigma}}, \quad (2.14)$$

which is extended to a vector valued stochastic variable as

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^{n_x} \det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}, \quad (2.15)$$

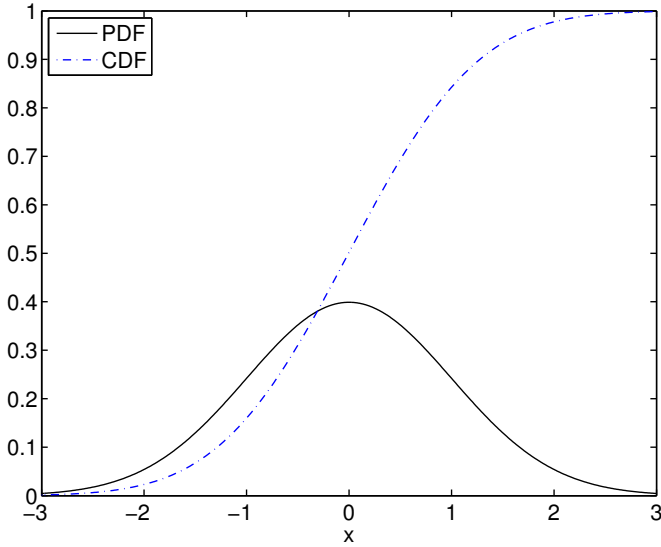
where  $n_x$  is the dimension of the vector  $x$ , see e.g., [31] for a more in detail description. The Gaussian CDF,

$$\Phi(x; \mu, \Sigma) := \int_{x' < x} \mathcal{N}(x', \mu, \Sigma) dx',$$

lacks an analytic expression, however, it is tabulated in most collection of statistical tables. Figure 2.1 shows the PDF and CDF of the normalized Gaussian distribution,  $\mathcal{N}(0, 1)$ .

The widespread usage of the Gaussian distribution is in part motivated by the fact that many natural phenomena exhibit Gaussian or Gaussian-like properties. One reason for this is that the sum of stochastic variables, under weak conditions by the *central limit theorem*, becomes Gaussian as the number of terms increases [68]. Hence, if a natural phenomenon is a combination of many stochastic phenomena it is often quite reasonable to assume that the combination of them is Gaussian.

The Gaussian distribution has favorable mathematical properties, and it is possible to derive many analytical results when the Gaussian distribution is used. One reason for this is that the Gaussian distribution is its own *conjugate prior*. (A conjugate prior is a



**Figure 2.1:** PDF and CDF for the normalized Gaussian distribution,  $\mathcal{N}(0, 1)$ .

prior chosen so that the posterior distribution is in the same family of distributions as the prior [72].) One such property is that any linear combination of Gaussian variables is Gaussian, *i.e.*, if  $x \sim \mathcal{N}(\mu, \Sigma)$  and  $B$  is a linear transformation of full (row) rank, then

$$z := Bx \sim \mathcal{N}(B\mu, B\Sigma B^T). \quad (2.16)$$

If  $B$  is rank deficient the resulting stochastic variable represents a degenerate case where one or more of the elements can be obtained as a combination of the other.

Finally, calculating the properties discussed in Section 2.1 for  $x \sim \mathcal{N}(\mu, \Sigma)$  yields  $\mathbb{E}(x) = \mu$ ,  $\text{cov}(x) = \Sigma$ ,  $\gamma_1 = 0$ , and  $\gamma_2 = 0$ . Furthermore, the intrinsic accuracy is  $\mathcal{I}_x = \Sigma^{-1}$  and subsequently Gaussian distributions have the relative accuracy  $\Psi_x = 1$ .

### 2.3.2 Gaussian Mixture Distribution

Even though the Gaussian distribution is common it is not powerful enough to capture every stochastic phenomena in a satisfactory manner. One way to extend the Gaussian distribution is to mix several Gaussian distributions. The result is a *Gaussian mixture distribution* or *Gaussian sum distribution*, defined by its PDF

$$\mathcal{N}_n(x; (\omega_\delta, \mu_\delta, \Sigma_\delta)_{\delta=1}^n) = \sum_{\delta=1}^n \omega_\delta \mathcal{N}(x; \mu_\delta, \Sigma_\delta), \quad (2.17)$$

where  $\omega_\delta > 0$ ,  $\sum_{\delta} \omega_\delta = 1$ , and  $n$  indicates how many Gaussian components are used. For  $n = 1$  the Gaussian mixture is Gaussian. Note, this notation is ambiguous, *e.g.*,  $\mathcal{N}_2((\frac{1}{2}, 0, 1), (\frac{1}{2}, 0, 1)) = \mathcal{N}(0, 1)$ . One interpretation of the Gaussian mixture is that

for all possible  $\delta$  the probability is  $\omega_\delta$  that a sample comes from the Gaussian distribution  $\mathcal{N}(\mu_\delta, \Sigma_\delta)$ . The result is a distribution that can be used to approximate any other distribution if  $n$  increases, [1, 2, 79]. The Gaussian mixture is also its own conjugate prior [72].

---

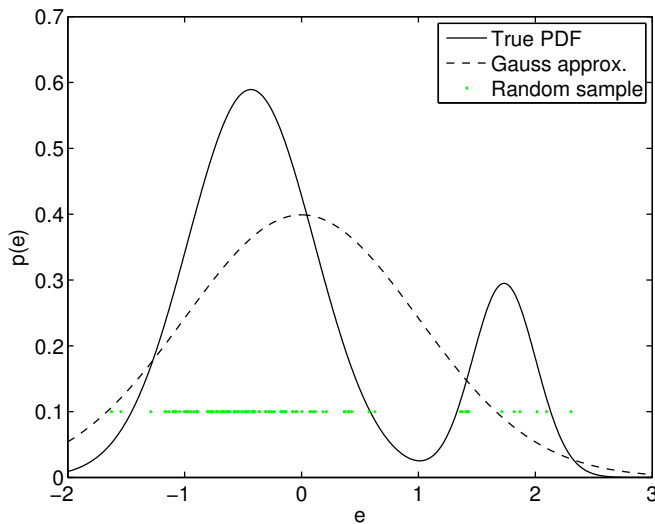
**Example 2.1: Bi-Gaussian noise**

---

Radar measurements often exhibit bimodal properties due to secondary radar reflections. The Master's theses [16, 81] both study this phenomenon for radar altitude measurements from an aircraft. The collected data clearly shows that measurements over certain terrains results in bimodal measurement errors. Radar reflections in treetops is one reason for this, e.g., when flying over forests. This noise,  $e$ , is well modeled using a Gaussian mixture with two components, a *bi-Gaussian* distribution, similar to

$$\begin{aligned} e &\sim \mathcal{N}_2((0.8, -0.43, 0.29), (0.2, 1.73, 0.07)) \\ &= 0.8\mathcal{N}(-0.43, 0.29) + 0.2\mathcal{N}(1.73, 0.07). \end{aligned} \quad (2.18)$$

The PDF of this distribution is depicted in Figure 2.2. Note that the Gaussian approximation poorly captures the true properties of the distribution. Hence, using knowledge of non-Gaussian noise characteristics can be favorable, as shown in e.g., [7–9].



**Figure 2.2:** The bimodal bi-Gaussian distribution in (2.18) and a second order equivalent Gaussian distribution. The distribution is representative of what can be found as measurement noise in a radar measurements.

---

The mean of a Gaussian mixture is obtained by a weighted sum of the means of the combined Gaussian components

$$\mu_x = \sum_{\delta} \omega_{\delta} \mu_{\delta}, \quad (2.19a)$$

and the covariance matrix a result of the contained covariances and spread of the mean factors

$$\Sigma_x = \sum_{\delta} \omega_{\delta} (\Sigma_{\delta} + \bar{\mu}_{\delta} \bar{\mu}_{\delta}^T), \quad (2.19b)$$

where  $\bar{\mu}_{\delta} := \mu_{\delta} - \mu_x$ . The higher order moments, assuming a scalar distribution, are

$$\gamma_{1,x} = \sum_{\delta} \omega_{\delta} \bar{\mu}_{\delta} (3\Sigma_{\delta} + \bar{\mu}_{\delta}) \Sigma_x^{-\frac{3}{2}} \quad (2.19c)$$

$$\gamma_{2,x} = \sum_{\delta} \omega_{\delta} (3\Sigma_{\delta}^2 + 6\bar{\mu}_{\delta}^2 \Sigma_{\delta} + \bar{\mu}_{\delta}^4) \Sigma_x^{-2} - 3. \quad (2.19d)$$

Compare these values to  $\gamma_1 = \gamma_2 = 0$  obtained for Gaussian distributions.

Calculating the intrinsic accuracy for a Gaussian mixture distribution is more difficult, and in general no analytic expression exists. However, Monte Carlo integration can be used to compute it. The gradient needed to compute the Fisher information *etc.*, is given by

$$\nabla_x \mathcal{N}_n(x) = \sum_{\delta=1}^n \omega_{\delta} \nabla_x \mathcal{N}(x; \mu_{\delta}, \Sigma_{\delta}) = - \sum_{\delta=1}^n \Sigma_{\delta}^{-1} (x - \mu_{\delta}) \mathcal{N}(x; \mu_{\delta}, \Sigma_{\delta}),$$

and samples from the Gaussian mixture can be obtained by randomly selecting a mode and then draw a sample from the Gaussian distribution indicated by the mode.

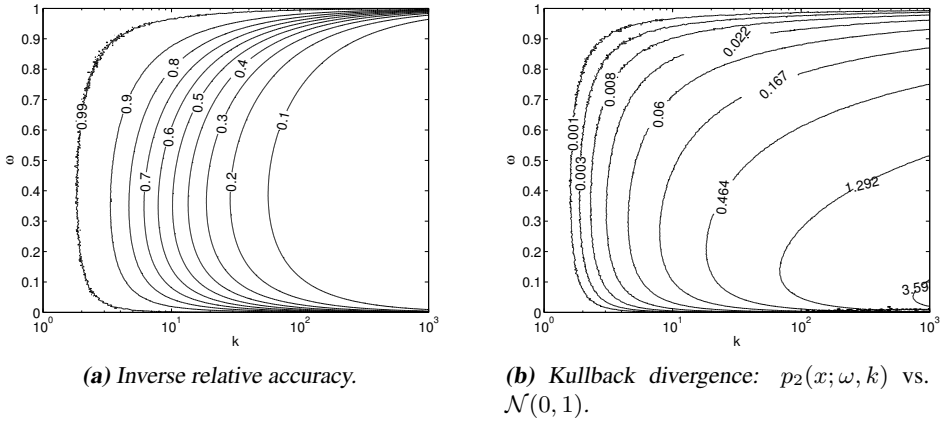
### Outliers Distribution

An example of a Gaussian mixture that will be used in this thesis to illustrate theoretical results is defined in terms of the PDF

$$p_1(x; \omega, k) = (1 - \omega) \mathcal{N}(x; 0, \Sigma) + \omega \mathcal{N}(x; 0, k\Sigma), \quad (2.20)$$

where  $\Sigma := 1/(1 + (k - 1)\omega)$ ,  $0 \leq \omega \leq 1$ , and  $k > 0$  to obtain proper distributions. The distribution can be used to model outliers. The two modes of the distribution represent nominal behavior and outlier behavior, respectively. With this interpretation outliers occur with probability  $\omega$  and have  $k$  times the nominal variance. Hence, if  $x \sim p_1(0.1, 10)$  then 10% of the samples are expected to be outliers and to have 10 times the variance of the nominal distribution.

The parameterization is intentionally chosen in such a way that  $E(x) = 0$  and  $\text{var}(x) = 1$  to allow for straightforward comparison with the normalized Gaussian distribution,  $\mathcal{N}(0, 1)$ . As with Gaussian distributions, moving the mean and changing the variance is a matter of changing the mean and scaling the variance of the different modes.



**Figure 2.3:** Inverse relative accuracy and Kullback divergence for the outliers description (2.20).

The relative accuracy for the parameterization (2.20) is given in Figure 2.3(a). To get the result Monte Carlo integration was used. Studying the contour plot shows that in an information perspective most information is available if about 30–40% of the samples are outliers, and the outliers have substantially larger variance.

The Kullback divergence between the outliers distributions,  $p_1(\omega, k)$ , and the normalized Gaussian distribution  $\mathcal{N}(0, 1)$ , has been computed to illustrate the difference between the distributions. The result is found in Figure 2.3(b). The relative accuracy and the Kullback divergence behaves similarly.

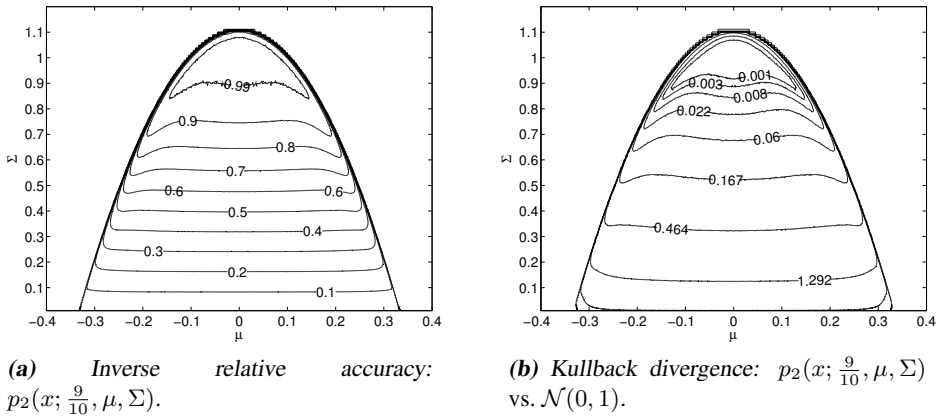
### Unsymmetric Bimodal Distribution

Another distribution that will be used throughout this thesis is the bimodal distribution given by the PDF

$$p_2(x; \omega, \mu, \Sigma) = \omega \mathcal{N}(x; \mu, \Sigma) + (1 - \omega) \mathcal{N}\left(x; \frac{-\omega\mu}{1-\omega}, \frac{1-\omega\mu^2/(1-\omega)-\omega\Sigma}{1-\omega}\right) \quad (2.21)$$

where  $0 < \Sigma < \frac{1}{\omega} - \frac{\mu^2}{1-\omega}$  and  $0 < \omega < 1$  to get a proper distribution. This is a distribution of the same type as was used in Example 2.1. As described there, this type of distributions can be used to model radar measurements, where one of the modes is the result of secondary radar reflections, e.g., in treetops [8, 16, 81]. However, multimodal distributions can be used to model other behaviors as well, for instance situations where there are two different possibilities where the means of the modes can be used to represent different behaviors.

The intrinsic accuracy of the parameterized distributions and the Kullback divergence compared to a normalized Gauss distribution are found in Figure 2.4, for  $\omega = \frac{9}{10}$ . The two measures behave very similarly, and in this case it is difficult to tell the two part. Close to the boundary of the allowed parameter region, the two modes both approach point distributions since the spread of the mean will contribute substantially to the total



**Figure 2.4:** Inverse relative accuracy and Kullback divergence for the bimodal distribution (2.21) with  $\omega = \frac{9}{10}$ .

variance of the distribution. With point distributions present the information content is very high, since once the point distribution is identified the estimate will be correct. A similar argumentation can be made about the Kullback information and its interpretation as a measure of how difficult it is to distinguish between two distributions.

### Symmetric Trimodal Distribution

The final distribution that will be used to illustrate the theory is defined by the PDF

$$p_3(x; \mu, \omega) = \frac{1-\omega}{2} \mathcal{N}(x; -\mu, \Sigma) + \omega \mathcal{N}(x; 0, \Sigma) + \frac{1-\omega}{2} \mathcal{N}(x; +\mu, \Sigma), \quad (2.22)$$

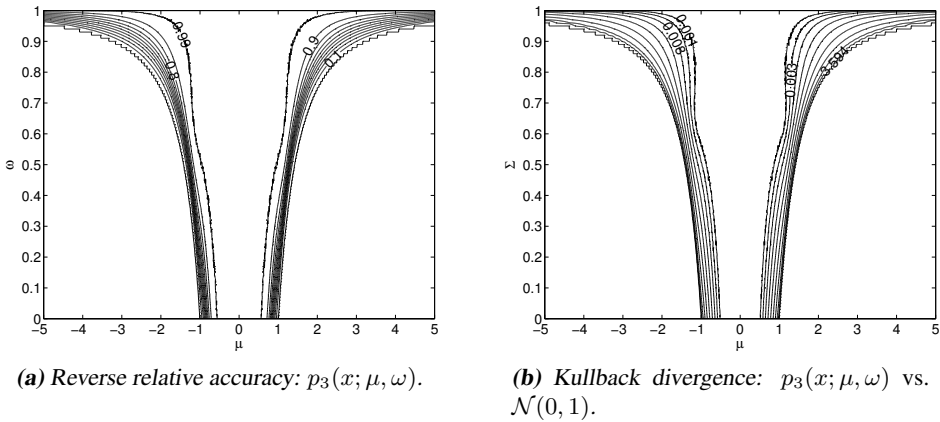
where  $\Sigma := 1 - \mu^2(1 - \omega)$  and  $1 - \mu^{-2} < \omega < 1$  to get a proper distribution. In many respects this distribution is similar to the bimodal distribution (2.21). One usage is to model different behaviors, where different means are interpreted as different behaviors that could be experienced.

The relative accuracy for (2.22) is found in Figure 2.5(a) and the Kullback divergence when the distribution is compared to a normalized Gaussian is in Figure 2.5(b). Once gain, the principal behavior of the relative accuracy and the Kullback divergence is similar. As the modes are separated the information increases up until the point where the distribution consists of three point distributions.

## 2.4 Transformed Distributions

At time to time a stochastic variable is passed through a transformation, e.g., as the result of a measurement. Assume that  $x$  is a stochastic variable passing through the function  $f(\cdot)$ , the result is a new stochastic variable  $z = f(x)$  with a new distribution.

As seen earlier, if  $x$  is Gaussian and  $f(\cdot)$  is a linear function then  $z$  is Gaussian as well, and the distribution is given by (2.16). The general way to determine the distribution of  $z$



**Figure 2.5:** Inverse relative accuracy and Kullback divergence for the trimodal distribution (2.22).

is to use the definition of a PDF,

$$p_z(z) = \frac{d}{dz} \int_{f(x) < z} p_x(x) dx, \tag{2.23}$$

which if  $f(\cdot)$  is bijective simplifies to

$$p_z(z) = \frac{d}{dz} \int_{f(x) < z} p_x(x) dx = \frac{d}{dz} \int_{-\infty}^{f^{-1}(z)} p_x(x) dx = p_x(z) \frac{df^{-1}(z)}{dz}. \tag{2.24}$$

The expressions above are given for scalar variables but can be extended to vector valued variables, see e.g., [31, Theorem 2.1 in Chapter I]

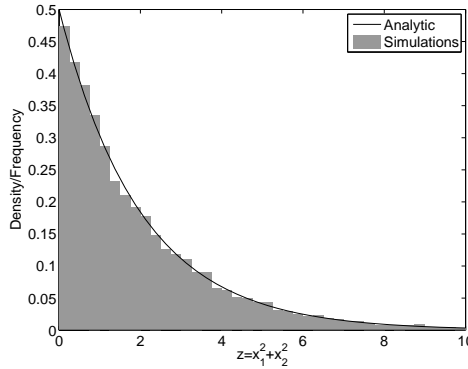
**Example 2.2:**  $\chi^2(2)$  distribution

Let  $x_1$  and  $x_2$  be two independent  $\mathcal{N}(0, 1)$  distributed stochastic variables, and let  $z = x_1^2 + x_2^2$ . Hence, if  $x_1$  and  $x_2$  are coordinates,  $z$  is the squared distance to origin. The distribution of  $z$  is then derived using (2.23),

$$\begin{aligned} p_z(z) &= \frac{d}{dz} \int_{x_1^2 + x_2^2 < z} \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-x_2^2/2} dx_1 dx_2 = \text{Change to polar coord.} / \\ &= \frac{d}{dz} \frac{1}{2\pi} \int_0^{\sqrt{z}} e^{-r^2/2} r dr \int_0^{2\pi} d\phi = \frac{d}{dz} \int_0^{\sqrt{z}} e^{-r^2/2} r dr = \frac{1}{2\sqrt{z}} e^{-z/2} \sqrt{z} = \frac{e^{-z/2}}{2}. \end{aligned}$$

The new distribution is  $\chi^2(2)$  distributed, and it can be shown that  $E(z) = 2$  and  $\text{var}(z) = 4$ . The  $\chi^2(2)$  distribution is a special instance of the  $\chi^2(n)$  distribution, the result obtained when adding  $n$  independent and squared  $\mathcal{N}(0, 1)$  stochastic variables.

The analytic  $\chi^2(2)$  PDF and the result of 1 000 simulated samples are shown in Figure 2.6. Computing the sample mean and variance yields results close to the analytic values.



**Figure 2.6:** Analytical and simulated (1 000 samples) PDF of  $\chi^2(2)$ .

Even though the target distribution in the example above turned out to be manageable to calculate, this is often not the case. Hence, approximations are needed. This section presents three such approximations: Monte Carlo transformation, linear Gaussian approximation, and unscented transformation.

### 2.4.1 Monte Carlo Transformation

If it is difficult to calculate (2.23) analytically, Monte Carlo integration can always be used to compute the integral (see Section 2.2). Using enough samples the approximation error can be made arbitrary small. Based on this, approximate the PDF with  $N$  samples according to

$$p_x(x) \approx \sum_{i=1}^N \omega^{(i)} \delta(x - x^{(i)}), \tag{2.25}$$

where  $\omega^{(i)} = 1/N$  if  $x^{(i)}$  are IID samples from  $x$  and  $\delta(\cdot)$  denotes the *generalized Dirac function*. Using some other distribution for  $x^{(i)}$  may be motivated, and then  $\omega^{(i)}$  follows according to Section 2.2.

To get an approximation of the target distribution transform all particles separately,

$$z^{(i)} = f(x^{(i)}),$$

and the PDF is then approximated by the sum

$$p_z(z) \approx \sum_{i=1}^N \omega^{(i)} \delta(z - z^{(i)}), \tag{2.26a}$$

where the combination of the importance weights and density of particles yields the PDF. Monte Carlo integration now yields the expected value and the covariance matrix:

$$\mu_z = \sum_{i=1}^N \omega^{(i)} z^{(i)} \quad (2.26b)$$

$$\Sigma_z = \sum_{i=1}^N \omega^{(i)} (z^{(i)} - \mu_z)(z^{(i)} - \mu_z)^T. \quad (2.26c)$$

The larger the set of particles, the better the approximation becomes.

## 2.4.2 Gauss Approximation Formula

The Gaussian approximation is to linearize the transformation around the expected value of  $x$  using a Taylor series expansion, and disregard all terms of order two or more. That is, the mean is transformed using,

$$\begin{aligned} \mu_z &= E_z(z) = E_x(f(x)) \\ &= E_x\left(f(\mu_x) + (\nabla_x f(\mu_x))^T(x - \mu_x) + \mathcal{O}(\|x - \mu_x\|^2)\right) \\ &\approx E\left(f(\mu_x)\right) + (\nabla_x f(\mu_x))^T E(x - \mu_x) = f(\mu_x), \end{aligned} \quad (2.27)$$

and the covariance matrix

$$\begin{aligned} \Sigma_z &= \text{cov}_z(z) = \text{cov}_x(f(x)) = E_x\left((f(x) - \mu_z)(f(x) - \mu_z)^T\right) \\ &= E_x\left(\left(f(\mu_x) + (\nabla_x f(\mu_x))^T(x - \mu_x) + \mathcal{O}(\|x - \mu_x\|^2) - f(\mu_x)\right)\right. \\ &\quad \cdot \left.\left(f(\mu_x) + (\nabla_x f(\mu_x))^T(x - \mu_x) + \mathcal{O}(\|x - \mu_x\|^2) - f(\mu_x)\right)^T\right) \\ &\approx (\nabla_x f(\mu_x))^T E_x((x - \mu_x)(x - \mu_x)^T) (\nabla_x f(\mu_x)) \\ &= (\nabla_x f(\mu_x))^T \Sigma_x (\nabla_x f(\mu_x)). \end{aligned} \quad (2.28)$$

Note that with the gradient notation used is  $\nabla_x(Ax) = A^T$ . The method gives an approximation of the first two moments of the distribution of  $z$ . Usually the distribution is then approximated with a Gaussian distribution with correct mean and covariance. This works fairly well in many situations where  $x$  is Gaussian and the transformation is fairly linear, but not always.

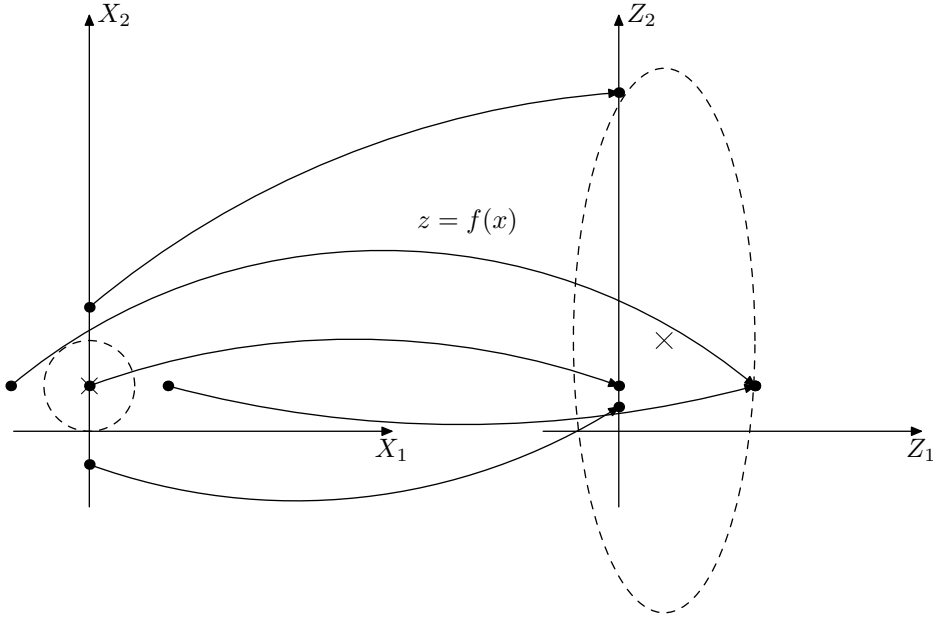
A straightforward extension to the Gaussian approximation is to include higher order terms in the approximation. The analysis is analogous, and since higher order moments of the transformation are captured the result is likely to be somewhat better.

## 2.4.3 Unscented Transform

The *unscented transform* (UT), introduced by Julier [38], Julier and Uhlmann [39, 40], is a recent method used to transform stochastic variables. The unscented transform was

designed as a step in the creation of the *unscented Kalman filter* (UKF). The UKF is discussed in Chapter 4 as an alternative filtering technique for nonlinear systems.

The basic idea of the unscented transform is to have a set of carefully chosen instances of the initial stochastic variable, called *sigma points*, pass through the transformation and based on these points derive the mean and the covariance matrix of the transformed distribution. The idea is illustrated in Figure 2.7.



**Figure 2.7:** Sigma points for  $x \sim \mathcal{N}((0, 1)^T, I)$  are passed through the nonlinear function  $z = f(x) = (x_1^2, x_2^2)^T$ . (A  $\times$  denote estimated mean and a dashed ellipse estimated covariance.)

The standard form of the unscented transform uses  $N = 2n_x + 1$  sigma points, where  $n_x$  is the dimension of  $x$ . The sigma points,  $x^{(i)}$ , and the associated weights,  $\omega^{(i)}$ , are chosen to be

$$x^{(0)} = E(x), \quad \omega^{(0)} \text{ is a parameter,} \quad (2.29a)$$

$$x^{(\pm i)} = x^{(0)} \pm \left[ \sqrt{\frac{n_x}{1 - \omega^{(0)}} \text{cov}(x)} \right]_i, \quad \omega^{(\pm i)} = \frac{1 - \omega^{(0)}}{2n_x}, \quad (2.29b)$$

for  $i = 1, \dots, n_x$ , where  $[A]_i$  denotes the  $i$ th column of  $A$ , and the square root is interpreted in the matrix sense  $A = \sqrt{A}\sqrt{A}$ , when necessary. The  $i$ th column of the square root of the covariance matrix is used since it represent the standard deviation of the distribution in a principal direction. The set of sigma points hence span the uncertainty of the stochastic variable. The weight on the mean,  $\omega^{(0)}$ , is used for tuning. A small value of  $\omega^{(0)}$  moves the sigma points away from the mean, whereas a large value gather them close to the mean. This allows for tuning of the unscented transform. The weight  $\omega^{(0)} = \frac{1}{3}$

gives, according to [40], preferable properties for Gaussian noise. It is also possible to use other sets of sigma points and/or parameterizations to change the behavior of the approximation and this way get more degrees of freedom for tuning [40].

Once the sigma points are chosen, the approximations of  $\mu_z$  and  $\Sigma_z$  are computed as weighted means. Denote the transformed sigma points

$$z^{(i)} = f(x^{(i)}),$$

for  $i = -n_x, \dots, n_x$  and associate them with the weights  $\omega^{(i)}$ . The estimated mean follows as

$$\mu_z = \sum_{i=-n_x}^{n_x} \omega^{(i)} z^{(i)}, \quad (2.30a)$$

and covariance matrix is

$$\Sigma_z = \sum_{i=-n_x}^{n_x} \omega^{(i)} (z^{(i)} - \mu_z)(z^{(i)} - \mu_z)^T. \quad (2.30b)$$

Once again, since only the mean and the covariance is known, a Gaussian approximation is often used to represent the result.

According to Julier and Uhlmann [40] it is possible to get correct estimates of mean and variance to the second order, and even higher orders, using the unscented transform. However, very little is said about how and under what conditions this holds, and Example 2.3 shows how the unscented transform sometimes produce poor approximations for relatively simple transformations.

### Example 2.3: Problem with unscented transform

Assume as in Example 2.2 two independent stochastic variables,  $x_1 \sim \mathcal{N}(0, 1)$  and  $x_2 \sim \mathcal{N}(0, 1)$ . The transformation  $z = x_1^2 + x_2^2$  is then  $\chi^2(2)$  distributed according to Example 2.2, with mean  $E(z) = 2$  and variance  $\text{var}(z) = 4$ .

Using the unscented transform with  $\omega^{(0)}$  as parameter yields and the sigma points

$$x^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad x^{(\pm 1)} = \begin{pmatrix} \pm \sqrt{\frac{2}{1-\omega^{(0)}}} \\ 0 \end{pmatrix}, \quad x^{(\pm 2)} = \begin{pmatrix} 0 \\ \pm \sqrt{\frac{2}{1-\omega^{(0)}}} \end{pmatrix},$$

Applying the transform results in the sigma points  $z^{(0)} = 0$  and  $z^{(\pm 1)} = z^{(\pm 2)} = \frac{2}{1-\omega^{(0)}}$  and the estimated mean and variance

$$\begin{aligned} \mu_z &= \sum_i \omega^{(i)} z^{(i)} = 2 \\ \Sigma_z &= \sum_i \omega^{(i)} (z^{(i)} - \mu_z)^2 = \frac{4\omega^{(0)}}{1-\omega^{(0)}}, \end{aligned}$$

where the variance differs from the variance for a  $\chi^2(2)$  distribution unless  $\omega^{(0)} = \frac{1}{2}$ . Hence, in this case the unscented transform, with  $\omega^{(0)}$  chosen as recommended in [40], does not produce an correct approximation of the distributions second moment, even for a quadratic function.

---

**Example 2.4: Comparison of transformation methods**


---

Let  $x \sim \mathcal{N}(1, 1)$  and  $z = x^2$ . The PDF for  $z$  is

$$p_z(z) = \frac{d}{dz} \int_{x^2 < z} \frac{1}{\sqrt{2\pi}} e^{-(x-1)^2/2} dx = \frac{1}{\sqrt{2\pi z}} e^{-(z+1)/2} \cosh(\sqrt{z}),$$

which turns out to be a non-central  $\chi^2$  distribution. Hence,  $z \sim \chi_1^2(1)$  with expected value  $E(z) = 2$  and variance  $\text{var}(z) = 6$ .

A Monte Carlo approximation of the target distribution using 1 000 particles yields  $\mu_z = 2.0$  and  $\Sigma_z = 6.0$ .

The Gauss approximation of the target distribution is

$$\begin{aligned} \mu_z &\approx \mu_x^2 = 1 \\ \Sigma_z &\approx 2\mu_x \cdot 1 \cdot 2\mu_x = 4 \end{aligned}$$

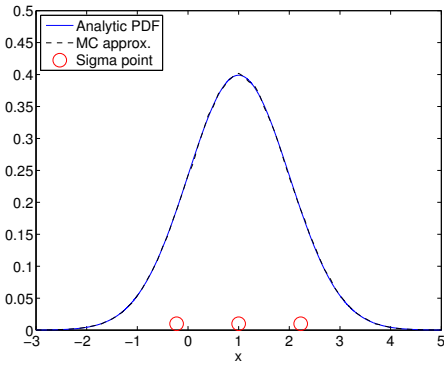
Finally, consider the unscented transform. The sigma points are using the weights  $\omega^{(0)} = \omega^{(\pm 1)} = \frac{1}{3}$

$$\begin{aligned} x^{(0)} &= \mu_x = 1, & z^{(0)} &= (x^{(0)})^2 = 1, \\ x^{(-1)} &= \mu_x - \sqrt{\frac{\text{var}(x)}{1 - \omega^{(0)}}} = 1 - \sqrt{\frac{3}{2}}, & z^{(-1)} &= (x^{(-1)})^2 = \frac{5}{2} - 2\sqrt{\frac{3}{2}}, \\ x^{(1)} &= \mu_x + \sqrt{\frac{\text{var}(x)}{1 - \omega^{(0)}}} = 1 + \sqrt{\frac{3}{2}}, & z^{(1)} &= (x^{(1)})^2 = \frac{5}{2} + 2\sqrt{\frac{3}{2}}, \end{aligned}$$

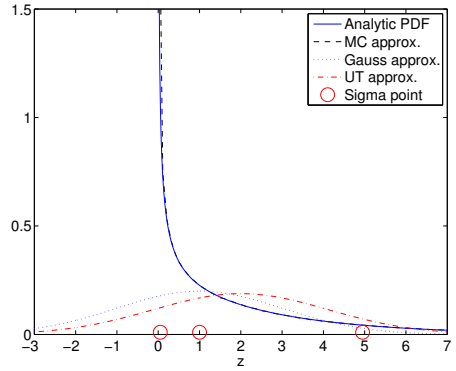
Using this

$$\begin{aligned} \mu_z &\approx \sum_{i=-1}^1 \omega^{(i)} z^{(i)} = 2 \\ \sigma_z &\approx \sum_{i=-1}^1 (z^{(i)} - \mu_z)^2 = \frac{9}{2} = 4.5. \end{aligned}$$

The results of applying the three approximative methods to derive the distribution of  $z$  is gathered in Figure 2.8. Note that performance is discouraging for the two the approximative methods, but the Monte Carlo transformation works well. The estimated distribution of  $z$  obtained with the Gauss approximation has both incorrect mean and incorrect variance. The unscented transform, with the suggested  $\omega^{(0)} = \frac{1}{3}$ , gets the mean correct, but underestimates variance. The PDF estimated in both cases differs significantly from the true distribution. Hence, it is important to make sure that the result is acceptable if an approximative transformation is used.



(a) PDF before the transformation.



(b) PDF after the transformation.

**Figure 2.8:** Result of squaring a normalized Gaussian distribution: analytically, with Gauss approximation formula, with Monte Carlo simulations and with unscented transformation. (Used sigma points are included for reference.)

# 3

---

## Models

**S**IR ISAAC NEWTON'S second law of motion (published in *Philosophiæ naturalis principia mathematica*, 1687) states: "The alternation of motion is proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed." In modern terms "the applied force equals the mass times the acceleration", which is a very precise description of how a force affects an object. This chapter deals with such descriptions of nature, in terms of mathematical *models*.

To extract information from measurements of a system, *i.e.*, estimate the underlying state of the system or detect a change in its behavior, it is important to have an accurate description of the system. A good system model describes how the measurable output reflects the internal state of the system and how it responds to input. For different applications, different descriptions are needed. They can be anything from textual descriptions, *e.g.*, "When the input voltage is 5 Volt, the output settles at 2 Volt.", or measured frequency response diagrams, or strict mathematical descriptions, *e.g.*, transfer functions or state-space models. Since the topic of this thesis is state estimation and change detection the emphasis of this section will be *discrete time models*; in particular system descriptions in terms of the similar *state-space model* and *hidden Markov model* (HMM).

### 3.1 State-Space Model

This section describes *state-space models*. The models given first are general in the sense that they can be used to describe a wide range of systems. Assumptions are then made that yield less general models, easier to analyze, but still powerful enough to be used in practice.

### 3.1.1 General State-Space Model

In a *state-space model*, a *state vector*,  $x_t$ , holds all information about the system, at time  $t$ , needed to determine its future behavior given the input. In this respect a state is very powerful since a low dimensional vector,  $x_t$ , can be used to summarize an infinite system history.

How the state evolves over time is determined by the system dynamics. The state at time  $t + 1$ ,  $x_{t+1}$ , relates to the state at time  $t$ ,  $x_t$ , (more generally times  $t_{i+1}$  and  $t_i$ ), and inputs of the system  $u_t$  and  $w_t$  through the relation

$$x_{t+1} = f(x_t, u_t, w_t), \quad (3.1a)$$

where  $u_t$  and  $w_t$  are both inputs but differ in their interpretation;  $u_t$  is a known input to the system, whereas  $w_t$ , usually referred to as *process noise*, is an unknown unpredictable input modeled as a stochastic process.

An observation of the system at time  $t$  is denoted  $y_t$ , and is a mapping of the state  $x_t$ , and the inputs  $u_t$  and  $e_t$ , where the latter is measurement noise, an unpredictable disturbance, modeled with a stochastic process. Generally, the measurement  $y_t$  can be related to  $x_t$ ,  $u_t$ , and  $w_t$  through the relation

$$0 = h(y_t, x_t, u_t, e_t). \quad (3.1b)$$

The functions  $f(\cdot)$  and  $h(\cdot)$  in (3.1) are both implicitly assumed to depend on the time  $t$  throughout this thesis, unless otherwise stated. One way to achieve this is to include  $t$  as an element in the state vector, or by explicitly adding time as an argument to the functions.

The model defined by (3.1) is very general. For simplicity it is often assumed that  $y_t$  can be solved for in (3.1b) and that  $e_t$  is additive. The result,

$$x_{t+1} = f(x_t, u_t, w_t) \quad (3.2a)$$

$$y_t = h(x_t, u_t) + e_t, \quad (3.2b)$$

is still a very general *nonlinear* model that is powerful enough to model almost any system encountered. In this thesis, (3.2) is the most general model used in the further discussions.

### 3.1.2 Linear State-Space Model

The state-space model (3.2) is good in the respect that it is general enough to describe almost any system. However, analysis of such models often becomes too complex to handle. This leads to the introduction of a new class of models, *linear models*. The general theory of linear systems is treated in e.g., [42, 74]. In linear models, the dynamics and the measurement relation are both considered to be linear functions of the state, according to

$$x_{t+1} = F_t x_t + G_t^u u_t + G_t^w w_t, \quad (3.3a)$$

$$y_t = H_t x_t + H_t^u u_t + e_t, \quad (3.3b)$$

where  $F_t$ ,  $G_t^u$ ,  $G_t^w$ ,  $H_t$ , and  $H_t^u$  are matrices.

---

**Example 3.1: Constant velocity model**


---

The *constant velocity* (CV) model describes a linear motion with constant velocity disturbed by external forces, process noise  $w_t$ , entering the system in terms of acceleration [61]. The constant velocity model is for instance used to track airplanes, where the deterministic but unknown maneuvers by the pilot can be treated as process noise.

In one dimension, using measurements of the position and a sampling time  $T$ , the model is given by

$$\begin{aligned}x_{t+1} &= \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix} x_t + \begin{pmatrix} \frac{1}{2}T^2 \\ T \end{pmatrix} w_t, \\y_t &= (1 \quad 0) x_t + e_t,\end{aligned}$$

where the states are position,  $x$ , and velocity,  $v$ , stacked as  $x = (x \quad v)^T$ . Furthermore,  $w_t$  and  $e_t$  must be described for a complete model. A common assumption is *white* noise (independent in time), Gaussian, and mutually independent noise, e.g.,  $w_t \sim \mathcal{N}(0, Q)$  and  $e_t \sim \mathcal{N}(0, R)$ .

To extend the model to handle multi-dimensional constant velocity motion, treat each dimension separately and stack the states describing the motion in each dimension in the state vector.

---

### 3.1.3 Model with Fault Terms

For applications such as fault detection it is preferable to introduce another input signal,  $f_t$ . With this modification  $f_t$  can be used to represent deterministic but to the magnitude unknown effects, e.g., faults such as a suddenly appearing friction force. The nonlinear model (3.2) with a fault term becomes

$$x_{t+1} = f(x_t, u_t, f_t, w_t), \quad (3.4a)$$

$$y_t = h(x_t, u_t, f_t) + e_t, \quad (3.4b)$$

and introducing the fault term in a linear model (3.3) yields

$$x_{t+1} = F_t x_t + G_t^u u_t + G_t^w w_t + G_t^f f_t, \quad (3.5a)$$

$$y_t = H_t x_t + H_t^u u_t + e_t + H_t^f f_t. \quad (3.5b)$$

Usually,  $f_t \equiv 0$  is assumed for the nominal system and any deviation from this is considered a fault or a change. Note that it is always possible to re-parameterize a given model in such a way that  $f_t$  is zero for the nominal model.

To make a distinction between regular noise, entering through  $w_t$  and  $e_t$ , and the effect of a fault or a change  $f_t$ , it will be assumed that  $f_t$  can be separated into a smoothly varying magnitude and a direction. Hence, the fault is parameterized as

$$f_t = H^{f,i} m_t, \quad (3.6)$$

where  $H^{f,i}$  gives the  $i$ th time-invariant fault direction with the scalar magnitude,  $m_t$ . The magnitude is then defined by the regression  $m_t = \varphi_t^T \theta$ , where  $\theta$  is constant over time and of relatively low dimension and where  $\phi_t$  makes up a basis for the fault behavior.

---

**Example 3.2: Incipient noise**


---

Consider an incipient fault with a magnitude that increases linearly in time. In the fault description introduced above, this behavior could be represented in the fault basis

$$\varphi_t = \begin{pmatrix} 1 \\ t \end{pmatrix} \quad \text{with the fault parameter} \quad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix},$$

where  $\theta_0$  captures the constant fault level and  $\theta_1$  determines the rate of the increase. Note that  $\theta$  is time-invariant even though the fault magnitude  $m_t = \varphi_t^T \theta$  is not.

To get a more advanced fault profile, include more terms to the basis of the magnitude, e.g., adding a quadratic term yields

$$\varphi_t = \begin{pmatrix} 1 \\ t \\ t^2 \end{pmatrix} \quad \text{and} \quad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix},$$

where  $\theta_2$  represents the quadratic term.

---

To make it easy to use the noise description, let  $\varphi_t$  define an orthonormal basis, such that  $\sum_{k=t-L+1}^t \varphi_k \varphi_k^T = I$ . One such suitable choice is the *discrete Chebyshev polynomials*. This definition preserves the fault energy, i.e.,  $\|m_t\|_2^2 = \sum_{k=t-L+1}^t m_k^2 = \|\theta\|_2^2$ . This method to distinguish between noise and unknown deterministic input is used in e.g., [35, 86].

### 3.1.4 Batched Linear State-Space Model

For applications where measurements are considered in batches it is convenient to have the model reflect this. Therefore, introduce the following notation for  $L$  stacked measurements,

$$\mathbb{Y}_t = \begin{pmatrix} y_{t-L+1} \\ y_{t-L+2} \\ \vdots \\ y_t \end{pmatrix}.$$

The same notation will be used throughout the thesis to represent other stacked variables, i.e.,  $\mathbb{X}_t$ ,  $\mathbb{W}_t$ ,  $\mathbb{E}_t$ ,  $\mathbb{U}_t$ , and  $\mathbb{F}_t$  will be used for  $L$  stacked  $x$ ,  $w$ ,  $e$ ,  $u$ , and  $f$ , respectively. Unless clearly stated otherwise,  $L$  is to be assumed to be such that the stacked vector includes all available data. Using stacked variables, it is possible to express the  $L$  measurements in a batch from the linear model (3.5) as

$$\mathbb{Y}_t = \mathcal{O}_t x_{t-L+1} + \bar{H}_t^w \mathbb{W}_t + \mathbb{E}_t + \bar{H}_t^u \mathbb{U}_t + \bar{H}_t^f \mathbb{F}_t, \quad (3.7)$$

where  $\mathcal{O}_t$  is the *extended observability matrix* that describes the impact of the initial state on the measurements, and  $\bar{H}_t^*$  matrices describing how  $\star \in \{w, u, f\}$  enters the measurements. A bar is used to separate these matrices from those used in the standard

linear model. The extended observability matrix is

$$\mathcal{O} = \begin{pmatrix} H_{t-L+1} \\ H_{t-L+2}F_{t-L+1} \\ \vdots \\ H_t \prod_{i=t-L+1}^{t-1} F_i \end{pmatrix} = \text{/if time-invariant/} = \begin{pmatrix} H \\ HF \\ \vdots \\ HF^{L-1} \end{pmatrix}, \quad (3.8a)$$

where an assumption about time-invariance simplifies the expression considerably. The input matrices are defined as

$$\begin{aligned} \bar{H}^* &= \begin{pmatrix} H_{t-L+1}^* & 0 & \dots & 0 \\ H_{t-L+2}G_{t-L+1}^* & H_{t-L+2}^* & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ H_t \prod_{i=t-L+1}^{t-1} F_i G_{t-L+1}^* & H_t \prod_{i=t-L+1}^{t-1} F_i G_{t-L+2}^* & \dots & H_t^* \end{pmatrix} \\ &= \text{/if time-invariant/} = \begin{pmatrix} H^* & 0 & \dots & 0 \\ HG^* & H^* & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ HF^{L-2}G^* & HF^{L-3}G^* & \dots & H^* \end{pmatrix}, \quad (3.8b) \end{aligned}$$

for  $\star \in \{w, u, f\}$ . Note, if the system is time-invariant the  $H^*$  are *Toeplitz matrices* which allows for improved numerical algorithms to be used. For a more complete description of this way to view the system see e.g., [28, 29, 86].

## 3.2 Hidden Markov Model

An alternative view of a system, is to study how probable different events or states are [72]. In Bayesian statistics, introduced by Bayes [6] in 1763, parameters are stochastic variables with distributions instead of unknown but deterministic values as in classical statistics. Prior knowledge and measurements are hence used to derive an as correct parameter distribution as possible, in this context focusing less on an exact parameter value. The *Hidden Markov model* (IMM) is a statistical model that has much in common with a state-space model. In an HMM, one Markov process represents the underlying system, which cannot be directly observed,

$$p(x_{t+1}|\mathbb{X}_t, \mathbb{U}_t) = p(x_{t+1}|x_t, u_t), \quad (3.9a)$$

with the initial information  $x_0 \sim p_0(\cdot)$ . Observations of the system are then made through a stochastic process, described by

$$p(y_t|\mathbb{X}_t, \mathbb{U}_t) = p(y_t|x_t, u_t). \quad (3.9b)$$

A tutorial on the subject of HMM can be found in [69].

The hidden Markov model has much in common with state-space models. In the HMM, the Markov process, which cannot be observed directly, corresponds to the description of the dynamics in the state space formulation, and the observations, described

with the stochastic process, relates to the measurement equation. Regular systems can be represented both as a HMM and as a state-space model. The process noise is then implicitly defined by the PDF (3.9a), for a linear model with  $G^{rw} = I$ , as

$$w_t \sim p(x_{t+1} - F_t x_t - G_t^u u_t \mid x_t, u_t).$$

The measurement noise is implicit in the PDF (3.9b), for additive noise, as

$$e_t \sim p(y_t - h(x_t, u_t) \mid x_t, u_t).$$

### Example 3.3: Hidden Markov model representation of a linear model

Consider a scalar linear model

$$\begin{aligned} x_{t+1} &= ax_t + w_t \\ y_t &= cx_t + e_t, \end{aligned}$$

where  $w_t \sim \mathcal{N}(0, Q)$  and  $e_t \sim \mathcal{N}(0, R)$ . The equivalent HMM is

$$\begin{aligned} p(x_{t+1} \mid x_t) &= \mathcal{N}(x_{t+1}; ax_t, Q), \\ p(y_t \mid x_t) &= \mathcal{N}(y_t; cx_t, R). \end{aligned}$$

Furthermore, for both models the distribution of  $x_0$  must be specified in order for the description to be complete, e.g.,  $x_0 \sim \mathcal{N}(0, \Pi_0)$ .

In estimation and detection, the inferred state distribution is very important, and with a HMM description of a system the inferred distribution of  $x_t$  follows directly from marginalization and Bayes' rule. Marginalization of (3.9a) with respect to the previous state yields the conditioned PDF

$$p(x_{t+1} \mid \mathbb{Y}_t) = \int p(x_{t+1} \mid x_t) p(x_t \mid \mathbb{Y}_t) dx_t, \quad (3.10a)$$

where the integration is performed over the whole state space of  $x_t$ . Information acquired from new measurements are incorporated into knowledge about the state using Bayes' rule,

$$p(x_t \mid \mathbb{Y}_t) = \frac{p(y_t \mid x_t) p(x_t \mid \mathbb{Y}_{t-1})}{\int p(y_t \mid x_t) p(x_t \mid \mathbb{Y}_{t-1}) dx_t} = \frac{p(y_t \mid x_t) p(x_t \mid \mathbb{Y}_{t-1})}{p(y_t \mid \mathbb{Y}_{t-1})}. \quad (3.10b)$$

The expressions (3.10) capture all information available about  $x_t$  given the measurements in  $\mathbb{Y}_t$  and the model. It is often hard, or impossible, to find closed form expressions for the inferred PDF of  $x_t$ . Especially the normalizing factor  $p(y_t \mid \mathbb{Y}_{t-1})$  is usually troublesome. Linear models with Gaussian noise is one class of models where an analytical solution is known, the Kalman filter [42]. The Kalman filter is further elaborated on in Chapter 4. In general *conjugate priors*, render manageable analytical results. The family of *exponential distributions*, of which the Gaussian distribution is a member, contains its own conjugate priors [72].

# 4

---

## Filtering Methods

**T**HE ART OF STATE ESTIMATION, is to learn as much about the state of a system as possible from measurements of it. This should be interpreted in the widest sense possible. The result is a time series of estimates, a *trajectory*, describing how the state has developed over time. If the objective is positioning a car to give the driver directions, the trajectory could be a list positions and speeds indicating where the car was and how fast it was going at certain times.

This chapter presents several different methods to utilize available sensor information for estimation purposes. First estimation in general is presented, and after that state estimation in particular. The first group of filtering methods introduced are the *sequential Monte Carlo methods*, or *particle filters*, which gives a general way to approximate (3.10) for the nonlinear system (3.2),

$$x_{t+1} = f(x_t, w_t) \tag{4.1a}$$

$$y_t = h(x_t) + e_t, \tag{4.1b}$$

where deterministic input  $u_t$  has been removed in favor of a clearer notation. The *Kalman filter* is presented next as an optimal and analytical solution to linear Gaussian estimation problem (3.3)

$$x_{t+1} = F_t x_t + G_t^w w_t, \tag{4.2a}$$

$$y_t = H_t x_t + e_t, \tag{4.2b}$$

where  $w_t$  and  $e_t$  are Gaussian, and as a suboptimal linear method for any other linear system. Finally, a few nonlinear variations of the Kalman filter are studied as well: the *extended Kalman filter* (EKF), the *unscented Kalman filter* (UKF), and *filter banks*.

## 4.1 Parameter Estimation

General parameter estimation techniques are covered thoroughly by most textbooks on estimation theory, e.g., [47, 58]. The aim of parameter estimation is to find  $\hat{x}$  that minimizes a certain *loss function*, or in a Bayesian setting a *cost function*,  $L(x, \mathbb{Y})$ ,

$$\hat{x} := \arg \min_{\hat{x}} L(\hat{x}, \mathbb{Y}),$$

where  $\mathbb{Y}$  are measurements related directly, or indirectly, to  $x$ . An estimate,  $\hat{x}$ , with expected value

$$E(\hat{x}) = x^0 + b,$$

where  $x^0$  is the true state, is *unbiased* if  $b = 0$ . If  $b \neq 0$  then the estimate bias is  $b$ . Unbiasedness is an important property for estimators in classical statistics, but of little interest in the Bayesian framework where the parameter distribution is the main objective and there is no true parameter value [72].

There are many estimation methods available, four important methods are listed below:

- *Least square* (LS) estimation, where the squared errors between the estimated and the actual measurements are minimized,

$$\hat{x}^{\text{LS}} := \arg \min_{\hat{x}} \sum_i \|y_i - h(\hat{x})\|_2^2. \quad (4.3a)$$

- *Maximum likelihood* (ML) estimation, where the estimate is chosen to be the parameter most likely to produce the measurements obtained,

$$\hat{x}^{\text{ML}} := \arg \max_{\hat{x}} p(\mathbb{Y}|\hat{x}). \quad (4.3b)$$

The ML estimate is optimal in the sense that it is unbiased and reaches the CRLB as the information in the measurements goes to infinity. However, the ML estimate is generally not unbiased for finite information.

- *Minimum mean square error* (MMSE) estimation, which basically is the Bayesian version of LS estimation,

$$\hat{x}^{\text{MMSE}} := \arg \min_{\hat{x}} E_{x, \mathbb{Y}}(x - \hat{x})^2 = E_{x|\mathbb{Y}}(x), \quad (4.3c)$$

where the last equality follows from performing the optimization.

- *Maximum a posteriori* (MAP) estimation, the Bayesian equivalent of ML, is defined as the most likely  $x$  based on the *a priori* distribution and the measurements,

$$\hat{x}^{\text{MAP}} := \arg \max_{\hat{x}} p(\hat{x}|\mathbb{Y}_t). \quad (4.3d)$$

In state estimation the estimate of a time-varying state,  $x_t$ , is sought and the estimate of  $x_t$ , using measurements up to time  $\tau$ ,  $\mathbb{Y}_\tau$ , is here denoted  $\hat{x}_{t|\tau}$ . If the PDFs (3.10),  $p(x_{t+1}|\mathbb{Y}_t)$  and  $p(x_{t+1}|\mathbb{X}_t)$ , are known estimation is easy just apply one of the estimators (4.3). However, most times it is impossible, or at least very difficult, to analytically derive these distributions.

## 4.2 Particle Filter

When an analytic solution is unavailable numeric approximations of (3.10) are necessary. The main problem with (3.10) is the integrals. One way to handle this is to grid the state space and approximate the PDFs with piecewise constant functions so that the integrals can be treated as sums that are easier to handle than integrals. This grid based approach, also called the *point-mass filter* (PMF), is described in depth in [14, 37, 53] and evaluated in [8, 52]. Another method, which has much in common with filter banks, is to use a Gaussian sum approximation [1, 79]. However, this method suffers heavily from the curse of dimensionality. In practice this drastically limits the usability of the point-mass filter. An alternative is to resort to stochastic methods, such as *Monte Carlo integration*, leading to *sequential Monte Carlo* filters, also known as *particle filters* (PF).

The foundation for the particle filter was laid in the 1950s [32], but the technique never came into use then. Reasons for this may include the lack of computing power needed to fully take advantage of the method, and that the fundamental resampling step, introduced by Gordon, Salmond, and Smith [27], was not yet used. Since the seminal paper [27] much has been written about the particle filter and its favorable properties, as well as applications using it. The particle filtering theory is described in [19, 20, 71], variations of the theory combining it with other filtering techniques in [51, 88], and applications in [9, 30, 45].

### 4.2.1 Approximative Probability Density Function

Based on the Monte Carlo integration technique and Monte Carlo transformation, described in Sections 2.2 and 2.4.1, a reasonable approximation of the interesting PDF is to use a set of  $N$  particles  $\{x_t^{(i)}\}_{i=1}^N$ , with associated weights  $\{\omega_{t|t-1}^{(i)}\}_{i=1}^N$ , such that

$$p(x_t | \mathbb{Y}_{t-1}) \approx \sum_{i=1}^N \omega_{t|t-1}^{(i)} \delta(x_t - x_t^{(i)}),$$

where the particles are IID samples from an importance distribution and the weights are matching importance weights. Properties such as the mean of  $x_t$ , the MSE, is then easily computed using Monte Carlo integration,

$$\hat{x}_{t|t} = \mathbb{E}_{x_t | \mathbb{Y}_{t-1}}(x_t) \approx \sum_{i=1}^N \omega_{t|t-1}^{(i)} x_t^{(i)}.$$

Now, assume that the measurement  $y_t$  is obtained, the updated PDF is then according to (3.10b)

$$p(x_t | \mathbb{Y}_t) \propto p(y_t | x_t) p(x_t | \mathbb{Y}_{t-1}),$$

leading to the importance weight update

$$\omega_{t|t}^{(i)} = \frac{p(y_t | x_t^{(i)}) \omega_{t|t-1}^{(i)}}{\sum_j p(y_t | x_t^{(j)}) \omega_{t|t-1}^{(j)}}.$$

The measurement updated PDF becomes, using the updated weights,

$$p(x_t | \mathbb{Y}_t) \approx \sum_{i=1}^N \omega_{t|t}^{(i)} \delta(x_t - x_t^{(i)}).$$

The following PDF must be constructed for the prediction, (3.10a),

$$p(x_{t+1} | \mathbb{Y}_t) = \int p(x_{t+1} | x_t) p(x_t | \mathbb{Y}_t) dx_t.$$

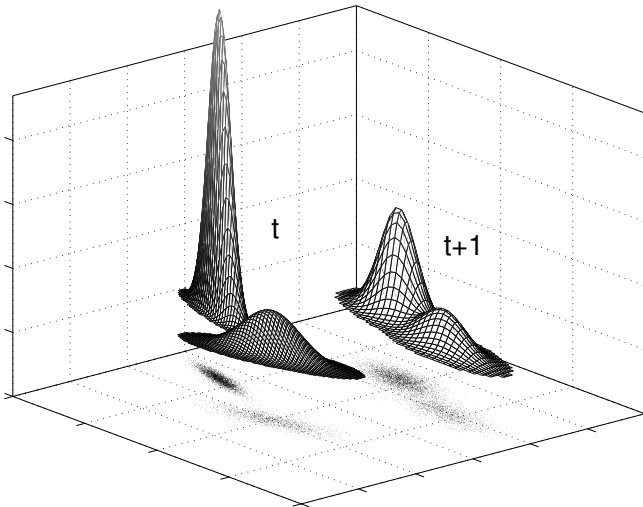
To do this, draw a new set of IID particles to represent the updated PDF, sample  $x_{t+1}^{(i)}$  from an importance distribution  $q(x_{t+1} | x_t^{(i)}, \mathbb{Y}_t)$  and update the importance weights accordingly,

$$\omega_{t+1|t}^{(i)} = \frac{p(x_{t+1}^{(i)} | x_t^{(i)})}{q(x_{t+1}^{(i)} | x_t^{(i)}, \mathbb{Y}_t)} \omega_{t|t}^{(i)},$$

yielding the approximative PDF

$$p(x_{t+1} | \mathbb{Y}_t) \approx \sum_{i=1}^N \omega_{t+1|t}^{(i)} \delta(x_{t+1} - x_{t+1}^{(i)}).$$

Figure 4.1 illustrates a particle cloud before and after a prediction step. Note how the cloud is translated by the transfer function  $f(\cdot, \cdot)$  and how the process noise spreads the particles. Compare this illustration to Figure 1.1 where the effects of a Gaussian approximation are shown.



**Figure 4.1:** PDFs for, and particles describing, the state distribution before (left) and after (right) the prediction step in a particle filter.

Note that if the importance distribution is chosen to be  $q(x_{t+1}|x_t, \mathbb{Y}_t) = p(x_{t+1}|x_t)$  this simplifies the update of the importance weights to  $\omega_{t+1|t}^{(i)} = \omega_{t|t}^{(i)}$ . However, this is not optimal with respect the statistical properties of the filter. Nevertheless, due to its simplicity this importance distribution is often used.

The above is a *sequential Monte Carlo method* and represents what was available in the 1950s [32]. It can be shown that using this method the approximated distribution will degenerate so that only a few particles actually contribute to the description of the PDF [19, 50]. Using  $q(x_{t+1}|x_t, \mathbb{Y}_t) = p(x_{t+1}|x_t)$  this can happen quite quickly, whereas a more careful choice of importance distribution may slow down the process. A solution to the problem is to introduce a *resampling* step as suggested in [27]. The resampling step in combination with the increased computational power was what was needed to turn the particle filter into an interesting method. Algorithm 4.1 represents a generic particle filter with resampling.

---

**Algorithm 4.1** Particle Filter
 

---

1. Initiate the filter:  $\{x_0^{(i)}\}_{i=1}^N \sim p(x_0)$  and  $\{\omega_{0|-1}^{(i)}\}_{i=1}^N = \frac{1}{N}$  for  $i = 1, \dots, N$ . Let  $t := 0$ .

2. Measurement update phase:

$$\omega_{t|t}^{(i)} = \frac{p(y_t|x_t^{(i)})\omega_t^{(i)}}{\sum_j p(y_t|x_t^{(j)})\omega_t^{(j)}}, \quad i = 1, \dots, N. \quad (4.4a)$$

3. Resample! See Section 4.2.2, e.g., Algorithms 4.2 and 4.3.

4. Time update phase: Generate  $N$  IID samples from an importance distribution  $q(\cdot)$ ,  $\{x_{t+1|t}^{(i)}\}_{i=1}^N \sim q(x_{t+1}|x_t^{(i)}, \mathbb{Y}_t)$  and update the importance weights accordingly,

$$\omega_{t+1|t}^{(i)} = \omega_{t|t}^{(i)} \frac{p(x_{t+1|t}^{(i)}|x_t^{(i)})}{q(x_{t+1|t}^{(i)}|x_t^{(i)}, \mathbb{Y}_t)}, \quad i = 1, \dots, N. \quad (4.4b)$$

If  $q(x_{t+1|t}|x_t^{(i)}, \mathbb{Y}_t) = p(x_{t+1|t}|x_t^{(i)})$  this simplifies to  $\omega_{t+1|t}^{(i)} = \omega_{t|t}^{(i)}$ .

5. Let  $t := t + 1$  and repeat from 2.

At any time in the algorithm a minimum variance estimate of the state can be obtained as a weighted sample mean:

$$\hat{x}_{t|t} = \sum_i \omega_{t|t}^{(i)} x_t^{(i)}, \quad \hat{x}_{t+1|t} = \sum_i \omega_{t+1|t}^{(i)} x_t^{(i)}.$$


---

## 4.2.2 Resampling

The resampling step rejuvenates the particles used to represent the PDF in the particle filter so that the stochastic support is maintained. Several methods have been devised to do this; one of the earliest and most commonly used resampling algorithms is the one used in the *sampling importance resampling* (SIR) particle filter [27] which is a special case of the more general *sequential importance sampling* (SIS) particle filter [19]. The SIR and SIS methods both use stochastic resampling. Other solutions to the problem rely on deterministic resampling. Four different resampling algorithms are compared and contrasted in [36]; two stochastic methods, and two deterministic methods. Resampling in a setting with a network of distributed computational nodes is reviewed in [13].

The SIR algorithm for resampling is straightforward and resampling is performed in each sample time. Resampling is done by drawing a new set of particles,  $x_{t+}^{(i)}$ , randomly with replacement from the old set of particles,  $x_t^{(i)}$ . The probability to choose a certain particle in the old set for the new set should be  $\Pr(x_{t+}^{(i)} = x_t^{(j)}) = \omega_{t|t}^{(j)}$ , and the weights matching the new particles are  $\omega_{t+|t}^{(i)} = 1/N$ . The new set of particles and weights then replace the old set in the particle filter. The SIR particle filter is given in Algorithm 4.2.

---

### Algorithm 4.2 Sampling Importance Resampling (SIR) Filter

---

Use Algorithm 4.1 with the following resampling step:

3. Resample: Construct the set  $\{x_{t+}^{(i)}\}_{i=1}^N$  by drawing  $N$  IID samples from  $\{x_t^{(j)}\}_{i=1}^N$ , with  $\Pr(x_{t+}^{(i)} = x_t^{(j)}) = \omega_{t|t}^{(j)}$ , and let  $\omega_{t+|t}^{(i)} = 1/N$  for  $i = 1, \dots, N$ . Let the new particles and weights replace the old ones:  $x_t^{(i)} := x_{t+}^{(i)}$  and  $\omega_{t|t}^{(i)} := \omega_{t+|t}^{(i)}$  for  $n = 1, \dots, N$ .
- 

The SIS algorithm is a generalization of the SIR algorithm, the difference lies in how often the resampling step is conducted. Whereas a SIR filter resamples before each time update, a SIS filter only resamples when the particle degeneration requires it. The *effective sample size* [8, 50],

$$N_{\text{eff}} = \frac{N}{1 + \text{var}_{q(x)}(\omega^{(i)})},$$

is one measure of the particle quality often used, but other measures have been suggested e.g., [80]. The effective sample size indicates how many of the particles that actually contribute to the support of the studied PDF. If  $N_{\text{eff}} \ll N$  this indicates that the support is poor and that resampling is needed to avoid degeneration of the filter. Unfortunately,  $N_{\text{eff}}$  is hard to calculate analytically, but can be approximated [50] with,

$$\hat{N}_{\text{eff}} = \frac{1}{\sum_i (\omega^{(i)})^2}.$$

Hence,  $\hat{N}_{\text{eff}}$  can be used to determine when to resample, *i.e.*, given some threshold  $N_{\text{th}}$  resample when  $\hat{N}_{\text{eff}} < N_{\text{th}}$ . One suggestion in [8] is to use  $N_{\text{th}} = \frac{2}{3}N$  as threshold for resampling. The SIS particle filter is given in Algorithm 4.3.

---

**Algorithm 4.3** Sequential Importance Sampling (SIS) Filter

---

Use Algorithm 4.1 with the following resampling step:

3. Resample: Compute the effective sample size,

$$\hat{N}_{\text{eff}} = \frac{1}{\sum_{i=1}^N (\omega_{t|t}^{(i)})^2}. \quad (4.5)$$

If  $\hat{N}_{\text{eff}} < N_{\text{th}}$  then construct the set  $\{x_{t+}^{(i)}\}_{i=1}^N$  by drawing  $N$  IID samples from  $\{x_t^{(j)}\}_{j=1}^N$ , with  $\Pr(x_{t+}^{(i)} = x_t^{(j)}) = \omega_{t|t}^{(j)}$ , and let  $\omega_{t+|t}^{(i)} = 1/N$  for  $i = 1, \dots, N$ .

Let the new particles and weights replace the old ones:  $x_t^{(i)} := x_{t+}^{(i)}$  and  $\omega_t^{(i)} := \omega_{t+}^{(i)}$  for  $i = 1, \dots, N$ .

---

## 4.3 Kalman Filter

Probably the most well-known estimation algorithm for linear systems is the *Kalman filter*. The filter is named after Rudolph E. Kalman, who in 1960 published a famous paper introducing the method [43]. At that time others were independently working with similar methods; amongst other Swerling [82] and researchers from the USSR [78]. Nevertheless Kalman has over time received most of the credit for developing the filter. In a few decades the Kalman filter became widely used. A collection of important work from the first decades of Kalman filtering, both theoretical and application oriented, can be found in [78]. The original Kalman filter assumes a discrete time linear model (3.3), but just a few years later the theory was extended to also include continuous time system [44].

The main idea of the Kalman filter is to use a linear filter to update the mean and the covariance of the estimate so that the covariance of the estimation error is kept minimal. If all noises are Gaussian, and hence stays Gaussian even after linear transformations, the Kalman filter proves the solution to (3.10) and yields the *minimum variance estimate* [42]. If the noise is non-Gaussian, the Kalman filter is the *best linear unbiased estimator* (BLUE), *i.e.*, it has the smallest error covariance of all linear filters, but there may exist nonlinear estimators that are better.

The Kalman filter is often derived for Gaussian noise where the calculations are straightforward to perform and the optimality follows immediately. There are many good books on the subject how to derive the Kalman filter, *e.g.*, the work by Anderson and Moore [2] and more lately the book by Kailath, Sayed, and Hassibi [42], therefore the filter is just given in Algorithm 4.4 without any derivation. The Kalman filter equations can be divided into a time update phase, where the dynamics of the system is handled, and a measurement update phase, where the measurements are incorporated in the estimate.

---

**Algorithm 4.4** Kalman Filter
 

---

1. Initiate the filter with the initial information:

$$\hat{x}_{0|-1} = x_0 \quad \text{and} \quad P_{0|-1} = \Pi_0,$$

where  $x_0$  is the initial state estimate and  $\Pi_0 = \text{cov}(x_0)$ . Let  $t = 0$ .

2. Measurement update phase:

$$K_t = P_{t|t-1} H_t^T (H_t P_{t|t-1} H_t^T + R_t)^{-1}$$

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t (y_t - H_t \hat{x}_{t|t-1} - H_t^u u_t) \quad (4.6a)$$

$$P_{t|t} = (I - K_t H_t) P_{t|t-1} \quad (4.6b)$$

3. Time update phase:

$$\hat{x}_{t+1|t} = F_t \hat{x}_{t|t} + G_t^u u_t \quad (4.6c)$$

$$P_{t+1|t} = F_t P_{t|t} F_t^T + Q_t \quad (4.6d)$$

4. Let  $t := t + 1$  and repeat from 2.
- 

Algorithm 4.4 is a straightforward way to implement a Kalman filter. However, numerical issues may be the result if the estimation problem is poorly stated. One such problem is that the covariance matrices lose symmetry. Another problem is that the covariances become indefinite. Symmetry is easily checked for, but indefiniteness is more difficult. One solution to this problem is to use a *square-root implementation*, also called an *array implementation*, of the Kalman filter algorithm where the square-root of the covariance matrix is propagated. This guarantees both symmetry and positive definiteness. Square-root based Kalman filter algorithms are described in e.g., [42]. The same technique is also used for other methods derived from the Kalman filter, e.g., the unscented Kalman filter discussed in Section 4.5.

## 4.4 Kalman Filters for Nonlinear Models

The Kalman filter, even though quite general, is in many cases not applicable due to its limitation to *linear models*. Many real life situations have either nonlinear dynamics or nonlinear measurements. Over time an approximative extension to the Kalman filter was developed, the *extended Kalman filter* (EKF) that uses linearized models to handle nonlinear models. Early examples of this are the works of Schmidt [75], Smith, Schmidt, and McGee [77]. An early reference where the EKF is thoroughly treated is Jazwinski [37]. The EKF is also treated in other standard references on filtering, e.g., [2, 42].

### 4.4.1 Linearized Kalman Filter

A first step towards a Kalman filter for nonlinear models is to assume that a nominal trajectory is known,  $x^{\text{nom}}$ , linearize the system about this trajectory (*cf.* Gaussian transformation of stochastic variables in Section 2.4.2) and then apply the Kalman filter to the linearized system. This is often called the *linearized Kalman filter* [42] and works in situations where a good nominal trajectory is known in advance, *e.g.*, the trajectory of an orbiting satellite.

The linear system resulting from linearization of (3.2) has the dynamics

$$\begin{aligned} x_{t+1} &= f(x_t, 0) \approx f(x_t^{\text{nom}}, 0) + F_t(x_t - x_t^{\text{nom}}) + G_t w_t \\ &= F_t x_t + \underbrace{f(x_t^{\text{nom}}, 0) - F_t x_t^{\text{nom}}}_{=: u'_t} + G_t w_t = F_t x_t + u'_t + G_t w_t, \end{aligned} \quad (4.7a)$$

without deterministic input for notational clarity. The gradient

$$F_t := \left( \nabla_x f(x, 0) \Big|_{x=x_t^{\text{nom}}} \right)^T, \quad G_t := \left( \nabla_w f(x_t^{\text{nom}}, w) \Big|_{w=0} \right)^T,$$

and the artificial input  $u'_t$  are all known at design time since the nominal trajectory  $x^{\text{nom}}$  is assumed known. Using the same technique the measurement relation can be expressed as

$$\begin{aligned} y_t &= h(x_t) + e_t \approx h(x_t^{\text{nom}}) + H_t(x_t - x_t^{\text{nom}}) + e_t \\ &= H_t x_t + \underbrace{h(x_t^{\text{nom}}) - H_t x_t^{\text{nom}}}_{=: u''_t} + e_t = H_t x_t + u''_t + e_t, \end{aligned} \quad (4.7b)$$

where the gradient

$$H_t := \left( \nabla_x h(x) \Big|_{x=x_t^{\text{nom}}} \right)^T.$$

and the artificial input  $u''_t$  are both known. It is now possible to use the linearized system (4.7) in the Kalman filter. Note that since the nominal trajectory is assumed known, it is possible to precompute  $P$  and  $K$ . However, if the nominal trajectory is not known, or not accurate enough to allow for the higher order terms in the linearization to be discarded, other methods are needed.

### 4.4.2 Extended Kalman Filter

In the *extended Kalman filter* (EKF) the problem with the lack of a nominal trajectory is solved using the information available from estimates by the filter itself. The best state estimate available, the latest estimate, can be used to construct a new linearization,

$$\begin{aligned} x_{t+1} &\approx f(\hat{x}_{t|t}, 0) + F_t(x_t - \hat{x}_{t|t}) + G_t w_t \\ &= F_t x_t + \underbrace{f(\hat{x}_{t|t}, 0) - F_t \hat{x}_{t|t}}_{=: u'_t} + G_t w_t = F_t x_t + u'_t + G_t w_t \end{aligned} \quad (4.8a)$$

$$\begin{aligned}
y_t &\approx h(\hat{x}_{t|t-1}) + H_t(x_t - \hat{x}_{t|t-1}) + e_t \\
&= H_t x_t + \underbrace{h(\hat{x}_{t|t-1}) - H_t \hat{x}_{t|t-1}}_{=: u_t''} + e_t = H_t x_t + u_t'' + e_t, \quad (4.8b)
\end{aligned}$$

with

$$F_t := \left( \nabla_x f(x, 0) \Big|_{x=\hat{x}_{t|t}} \right)^T, \quad G_t := \left( \nabla_w f(\hat{x}_{t|t}, w) \Big|_{w=0} \right)^T,$$

and

$$H_t := \left( \nabla_x h(x) \Big|_{x=\hat{x}_{t|t-1}} \right)^T.$$

For this linearization,  $K$  and  $P$  cannot be computed at design time, because neither  $\hat{x}_{t|t}$  nor  $\hat{x}_{t|t-1}$  are available beforehand, both become available before they are needed in the recursion. The result, found in Algorithm 4.5, is the EKF. This filter is harder to analyze than the linearized Kalman filter, but in practice it has proven to work well in many applications.

---

#### Algorithm 4.5 Extended Kalman Filter

---

1. Initiate the filter with the initial information:

$$\hat{x}_{0|-1} = x_0 \quad \text{and} \quad P_{0|-1} = \Pi_0,$$

where  $x_0$  is the initial state estimate and  $\Pi_0 = \text{cov}(x_0)$ . Let  $t = 0$ .

2. Measurement update phase:

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(y_t - h(\hat{x}_{t|t-1})) \quad (4.9a)$$

$$P_{t|t} = (I - K_t H_t) P_{t|t-1} \quad (4.9b)$$

$$K_t = P_{t|t-1} H_t^T (H_t P_{t|t-1} H_t^T + R_t)^{-1}$$

with

$$H_t := \left( \nabla_x h(x) \Big|_{x=\hat{x}_{t|t-1}} \right)^T.$$

3. Time update phase:

$$\hat{x}_{t+1|t} = f(\hat{x}_{t|t}, 0) \quad (4.9c)$$

$$P_{t+1|t} = F_t P_{t|t} F_t^T + G_t Q_t G_t^T \quad (4.9d)$$

where

$$F_t := \left( \nabla_x f(x, 0) \Big|_{x=\hat{x}_{t|t}} \right)^T \quad \text{and} \quad G_t := \left( \nabla_w f(\hat{x}_{t|t}, w) \Big|_{w=0} \right)^T$$

4. Let  $t := t + 1$  and repeat from 2.
-

### 4.4.3 Iterated Extended Kalman Filter

One problem that can sometimes occur with the EKF is that if the estimate is poor, then the filter gain matrix  $K$  is affected due to linearizing around a point far from the true state. This is most serious when the functions involved are highly nonlinear. The *iterated extended Kalman filter* (IEKF) is a slight modification of the EKF aimed to improve this [37, 42].

The main idea in the IEKF is that the measurement update most likely improves the state estimate, and that if the new estimate is used for the linearization this should improve performance further. To utilize this idea, the measurement update of the EKF is performed repeatedly using what is likely to be better and better estimates of the state as they become available. This is obtained by replacing the measurement update (4.9a)–(4.9b) in Algorithm 4.5 with

$$\hat{x}_{t|t} = \hat{x}_{t|t}^{(m)} \quad (4.10a)$$

$$P_{t|t} = (I - K_t^n H_t^{(m)}) P_{t|t-1}, \quad (4.10b)$$

where  $\hat{x}_{t|t}^{(m)}$ ,  $K_t^{(m)}$ , and  $H_t^{(m)}$  are given by the recursion

$$\begin{aligned} \hat{x}_{t|t}^{(0)} &= \hat{x}_{t|t-1} \\ \hat{x}_{t|t}^{(i)} &= \hat{x}_{t|t-1} + K_t^{(i)} (y_t - h(\hat{x}_{t|t-1})) \\ K_t^{(i)} &= P_{t|t-1} H_t^{(i)T} (H_t^{(i)} P_{t|t-1} H_t^{(i)T} + R_t)^{-1} \\ H_t^{(i)} &:= \left( \nabla_x h(x) \Big|_{x=\hat{x}_{t|t-1}^{(i)}} \right)^T. \end{aligned}$$

The number of iterative improvements performed is determined by the parameter  $m$ . Choosing  $m$  to be 3–5 is usually enough to obtain the desired effect.

## 4.5 Unscented Kalman Filter

The EKF is sufficient for many applications, if nothing else the popularity and wide spread usage of the EKF is a proof of this. However, there are situations when the EKF performs poorly. Furthermore, to use an EKF gradients must be computed and computing the gradients can be computationally expensive. Therefore, a scheme, not needing any gradients and which avoids numerical problems such as indefinite and unsymmetric covariance matrices by not performing the Riccati recursion explicitly, based on the *unscented transform* discussed in Section 2.4.3, has been suggested by Julier [38] and Julier and Uhlmann [39, 40]. This *unscented Kalman filter* (UKF) and aspects of it has also been discussed in for instance [87, 91, 92]. The UKF and other similar approaches can be gathered in a common framework of *linear regression Kalman filters*, [56, 57].

The basic idea of the UKF, and other linear regression Kalman filters, is to use a set of carefully chosen points in the state space to capture the effect of model nonlinearities on means and covariances during filtration. The UKF uses the *unscented transform* [39] to

select points from an augmented state space that includes the state,  $x_t$ , the process noise,  $w_t$ , and the measurement noise,  $e_t$ , in one state vector,

$$\mathcal{X}_t = \begin{pmatrix} x_t \\ w_t \\ e_t \end{pmatrix},$$

with the dimension  $n_{\mathcal{X}}$ . Using the augmented state vector  $\mathcal{X}_t$  allows for a straightforward utilization of the unscented transform discussed in Section 2.4.3, especially when it comes to find suitable sigma points. However, once the sigma points have been acquired the augmented state can be split up again to give a more familiar notation. By letting the sigma points pass through the model dynamics the following time update phase is obtained,

$$x_{t+1|t}^{(i)} = f(x_{t|t}^{(i)}, w_t^{(i)}),$$

and the result can then be combined to obtain the estimate

$$\begin{aligned} \hat{x}_{t+1|t} &= \sum_{i=-n_{\mathcal{X}}}^{n_{\mathcal{X}}} \omega_t^{(i)} x_{t+1|t}^{(i)} \\ P_{t+1|t} &= \sum_{i=-n_{\mathcal{X}}}^{n_{\mathcal{X}}} \omega_t^{(i)} (x_{t+1|t}^{(i)} - \hat{x}_{t+1|t}) (x_{t+1|t}^{(i)} - \hat{x}_{t+1|t})^T. \end{aligned}$$

The information in the measurements is introduced in a similar way by first obtaining the predicted measurements based on the sigma points

$$y_t^{(i)} = h(x_t^{(i)}, e_t^{(i)}),$$

yielding

$$\begin{aligned} \hat{y}_t &= \sum_{i=-n_{\mathcal{X}}}^{n_{\mathcal{X}}} \omega_t^{(i)} y_t^{(i)} \\ S_t &= \sum_{i=-n_{\mathcal{X}}}^{n_{\mathcal{X}}} \omega_t^{(i)} (y_t^{(i)} - \hat{y}_t) (y_t^{(i)} - \hat{y}_t)^T. \end{aligned}$$

The filter gain,  $K_t$ , is then computed as the cross-covariance between state and measurement divided by the covariance of the state, as this will project the state onto the new innovation,

$$K_t = \sum_{i=-n_{\mathcal{X}}}^{n_{\mathcal{X}}} \omega_t^{(i)} (x_{t|t-1}^{(i)} - \hat{x}_{t|t-1}) (y_t^{(i)} - \hat{y}_t)^T S_t^{-1},$$

and the estimate follows as

$$\begin{aligned} \hat{x}_{t|t} &= \hat{x}_{t|t-1} + K_t (y_t - \hat{y}_t) \\ P_{t|t} &= P_{t|t-1} - K_t S_t K_t^T. \end{aligned}$$

The UKF is given in detail in Algorithm 4.6.

**Algorithm 4.6** Unscented Kalman Filter

1. Initiate the filter with the initial information:

$$\hat{x}_{0|0} = x_0 \quad \text{and} \quad P_{0|0} = \Pi_0,$$

where  $x_0$  is the initial estimate and  $\Pi_0 = \text{cov}(x_0)$ . Let  $t = 1$ .

2. Choose  $N$  sigma points,  $\mathcal{X}_t^{(i)}$  partitioned as  $\mathcal{X}_t = \begin{pmatrix} x_t \\ w_t \\ e_t \end{pmatrix}$ , and weights,  $\omega_t^{(i)}$ , e.g., as suggested by (2.29).
3. Time update phase:

$$\hat{x}_{t|t-1} = \sum_{i=0}^N \omega_t^{(i)} x_{t|t-1}^{(i)} \quad (4.11a)$$

$$P_{t+1|t} = \sum_{i=0}^N \omega_t^{(i)} (x_{t|t-1}^{(i)} - \hat{x}_{t|t-1})(x_{t|t-1}^{(i)} - \hat{x}_{t|t-1})^T \quad (4.11b)$$

where

$$x_{t|t-1}^{(i)} = f(x_{t-1|t-1}^{(i)}, w_t^{(i)}).$$

4. Measurement update phase:

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t (y_t - \hat{y}_t) \quad (4.11c)$$

$$P_{t|t} = P_{t|t-1} - K_t S_t K_t^T, \quad (4.11d)$$

where

$$y_t^{(i)} = h(x_{t|t-1}^{(i)}, e_t^{(i)})$$

$$\hat{y}_t = \sum_{i=0}^N \omega_t^{(i)} y_t^{(i)}$$

$$S_t = \sum_{i=0}^N \omega_t^{(i)} (y_t^{(i)} - \hat{y}_t)(y_t^{(i)} - \hat{y}_t)^T$$

$$K_t = \sum_{i=0}^N \omega_t^{(i)} (x_{t|t-1}^{(i)} - \hat{x}_{t|t-1})(y_t^{(i)} - \hat{y}_t)^T S_t^{-1}.$$

5. Let  $t := t + 1$  and repeat from 2.

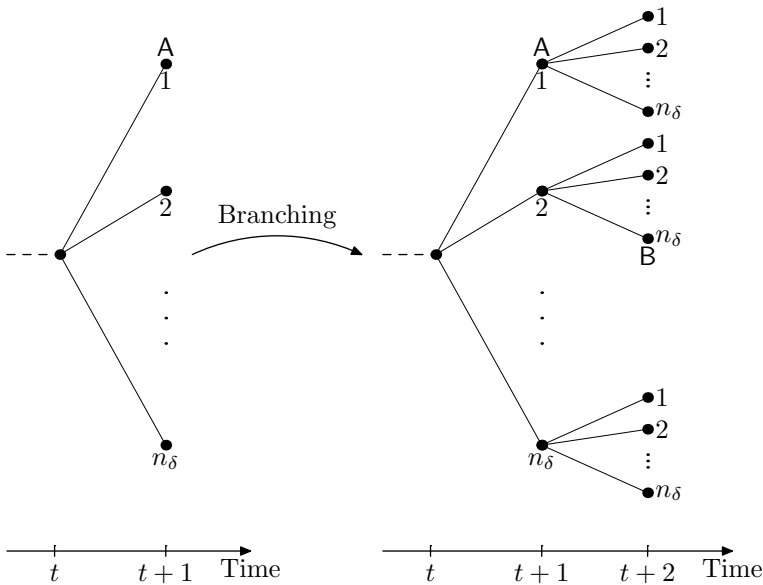
## 4.6 Filter Banks

It is sometimes useful to have models that behave differently depending on a discrete mode, e.g., an operational mode that changes the model characteristic. To track a highly maneuverable target one mode can be designed to track small maneuvers and one to handle more advanced maneuvers. This can be described as a special case of (3.1),

$$x_{t+1} = f(x_t, w_t, \delta_t) \quad (4.12a)$$

$$y_t = h(x_t, e_t, \delta_t), \quad (4.12b)$$

where  $\delta_t$  is used to indicate the mode at time  $t$  and  $p_{\delta|\delta'}$  is the probability that the model changes from mode  $\delta'$  to  $\delta$ . Even though not explicitly stated, it should be assumed that  $w_t$  and  $e_t$  could both be affected by the active mode. This way the method can be used to approximate any process noise and measurement noise arbitrarily well [1, 79]. How the system evolves over time is determined by which modes it has visited. Hence, introduce  $\delta_t$ , in analogy with  $\mathbb{Y}_t$ , to hold the mode history. If there are  $n_\delta$  different possible modes at each sample time this yields  $n_\delta^L$  possible different mode combinations to keep track of over a window of  $L$  samples. Thus, the mode complexity increases exponentially over time. The basic branching idea and the exponentially increasing number of nodes are illustrated in Figure 4.2.



**Figure 4.2:** Illustration of the exponential increase in mode combinations, i.e., how the model branch off new modes in each time step. Each node represents one possible mode combination. The nodes marked A indicate  $\delta_{t+1} = (\delta_t, \delta_{t+1} = 1)$  and the B node  $\delta_{t+2} = (\delta_t, \delta_{t+1} = 2, \delta_{t+2} = n_\delta)$ .

The structure of a *multiple models* (MM) model such as (4.12) can be utilized for filter construction. The idea is to treat each mode of the model independently, and combine

the independent results based on the likelihood of the obtained measurements. The number of parallel filters must in practice be kept from increasing exponentially and different approaches have been developed to achieve this. Two major and closely related methods are the *generalized pseudo-Bayesian* (GPB) filter [11] and the *multiple interacting models* (IMM) filter [11]. Especially the latter is popular in the tracking community. The *adaptive forgetting through multiple models* (AFMM) by Andersson [3] is another solution to the problem. The three different filters differ mainly in how and when the exponential complexity is avoided, and in how the estimate is computed. This section covers how to maintain a perfect multiple model filter bank, and then a *pruning* algorithm, related to the AFMM, is used to illustrate a way to handle the growing filter complexity.

### 4.6.1 Complete Filter Bank

To get the inferred PDF for (4.12) one filter for each possible mode is maintained and associated with a probabilities based on the model and available measurements. Given that there are filters to correctly handle the different modes of the model, the result is optimal in the sense that it produce the correct *a posteriori* distribution for  $x_t$ .

To do this a *filter bank*,  $\mathcal{F}$ , is maintained to hold the filters and their probability. If Kalman filters, or EKFs, are used, the filter bank  $\mathcal{F}_{t|t}$  reduces to a set of quadruples  $(\delta_t, \hat{x}_{t|t}^{(\delta_t)}, P_{t|t}^{(\delta_t)}, \omega_{t|t}^{(\delta_t)})$  representing mode, estimate, covariance matrix, and probability of the filter operating in that mode.

Now, assume that a filter bank,  $\mathcal{F}_{t|t}$ , is available that incorporates all information available from the model and measurements at time  $t$ . That is, there is one entity in it for every node at the  $t$  level of the tree in Figure 4.2.

In order for the filter bank to evolve in time so that it correctly represents the *posterior* state distribution, it must now *branch*, i.e., for each filter in  $\mathcal{F}_{t|t}$   $n_\delta$  new filters should be created, one for each possible mode at the next time step. These new filters obtain their internal state from the filter they are derived from and are then time updated. The probability of these new filters are

$$\omega_{t+1|t}^{(\delta_{t+1})} = p(\delta_{t+1}|\mathbb{Y}) = p(\delta_{t+1}|\delta_t)p(\delta_t|\mathbb{Y}_t) = p_{\delta_{t+1}|\delta_t}\omega_{t|t}^{(\delta_t)}.$$

The new filters together with the associated probabilities make up the filter bank  $\mathcal{F}_{t+1|t}$ .

The next step is to update the filter bank when a measurement arrives. This too is done in two steps. First, each individual filter in  $\mathcal{F}_{t|t-1}$  is updated using standard measurement update methods, e.g., a Kalman filter, and then the probability is updated according to how probable that mode is given the measurement,

$$\omega_{t|t}^{(\delta_t)} = p(\delta_t|\mathbb{Y}_t) = \frac{p(y_t|\delta_t, \mathbb{Y}_{t-1})p(\delta_t|\mathbb{Y}_{t-1})}{p(y_t|\mathbb{Y}_t)} = \frac{p(y_t|\delta_t)\omega_{t|t-1}^{(\delta_t)}}{\sum_{\delta_t} p(y_t|\delta_t)\omega_{t|t-1}^{(\delta_t)}},$$

where the last step requires that the measurements are mutually independent and that only the present state affects the measurement. The filters and the associated weights can now be gathered in the filter bank  $\mathcal{F}_{t|t}$ .

An algorithm to create a complete filter bank is given in Algorithm 4.7.

---

**Algorithm 4.7** Filter Bank Algorithm
 

---

1. Let  $\mathcal{F}_{0|-1}$  representing the initial state knowledge, e.g., with one single entry  $(\delta_0, x_0, \Pi_0, 1)$ . Let  $t = 0$ .
2. Measurement update all filters in  $\mathcal{F}_{t|t-1}$  and put the result in  $\mathcal{F}_{t|t}$ . Update the mode probabilities in  $\mathcal{F}_{t|t}$  using

$$\omega_{t|t}^{(\delta_t)} = \frac{p(y_t|\delta_t)\omega_{t|t-1}^{(\delta_t)}}{\sum_{\delta_t} p(y_t|\delta_t)\omega_{t|t-1}^{(\delta_t)}} \quad (4.13a)$$

The minimum variance estimate based on this is:

$$\hat{x}_{t|t} = \sum_{\delta_t} \omega_{t|t}^{(\delta_t)} \hat{x}_{t|t}^{(\delta_t)} \quad (4.13b)$$

$$P_{t|t} = \sum_{\delta_t} \omega_{t|t}^{(\delta_t)} \left( P_{t|t}^{(\delta_t)} + (\hat{x}_{t|t}^{(\delta_t)} - \hat{x}_{t|t})(\hat{x}_{t|t}^{(\delta_t)} - \hat{x}_{t|t})^T \right). \quad (4.13c)$$

3. Branch the filter bank, i.e., from  $\mathcal{F}_{t|t}$  construct  $\mathcal{F}_{t+1|t}$  by copying each entry in  $\mathcal{F}_{t|t}$   $n_\delta$  times to the new filter bank only updating the current mode, and the probability of that mode. The probability of the modes are updated using

$$\omega_{t+1|t}^{(\delta_{t+1})} = p_{\delta_{t+1}|\delta_t} \omega_{t|t}^{(\delta_t)}. \quad (4.13d)$$

The minimum variance estimate based on this is:

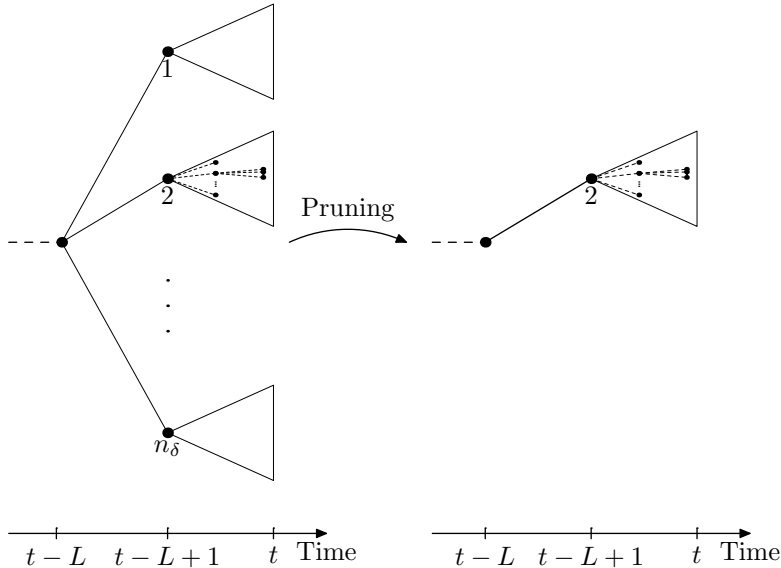
$$\hat{x}_{t+1|t} = \sum_{\delta_{t+1}} \omega_{t+1|t}^{(\delta_{t+1})} \hat{x}_{t+1|t}^{(\delta_{t+1})} \quad (4.13e)$$

$$P_{t+1|t} = \sum_{\delta_{t+1}} \omega_{t+1|t}^{(\delta_{t+1})} \left( P_{t+1|t}^{(\delta_{t+1})} + (\hat{x}_{t+1|t}^{(\delta_{t+1})} - \hat{x}_{t+1|t})(\hat{x}_{t+1|t}^{(\delta_{t+1})} - \hat{x}_{t+1|t})^T \right). \quad (4.13f)$$

4. Time update all filters in  $\mathcal{F}_{t+1|t}$ .
  5. Let  $t := t + 1$  and repeat from 2.
- 

## 4.6.2 Filter Bank with Pruning

The method used in this section demonstrates how to reduce the size of a filter bank by simply throwing away, *pruning*, unlikely mode combinations to allow for getting focus on more probable ones. To achieve this only minor modifications to Algorithm 4.7 are needed.



**Figure 4.3:** Illustration of pruning of a filter bank represented as a tree. In the left tree, before pruning, the subtree (represented by a triangle) originating at  $\delta_{t-L+1} = 2$  is the most probable and the other modes are removed. The result is displayed to the right.

To keep a complete filter bank over a window of  $L$  samples the filter bank must hold at least  $(n_\delta)^L$  filters with associated probabilities. This way the window length  $L$  becomes a parameter for tuning; if  $L$  is large the inferred PDF is accurate at the cost of a large filter bank, if  $L$  is small the filter bank is small but the approximation less accurate. A rule of thumb is to choose  $L \geq n_x + 1$  so that the modes get a chance to manifest themselves before being removed [28].

The reduction of the filter bank, the pruning, is performed the following way. Since a complete mode history over a window of  $L$  samples are asked for, the different modes at time  $T - L + 1$  is examined. This should be the first time instance where multiple modes are present. Determine which mode at this level that is most probable and remove all filters in the filter bank not stemming from this mode. This reduces the filter bank to one  $n_\delta$ th of the size. Hopefully the removed filters carry little weight, hence affecting the estimate little. As a last step renormalize the weights. The pruning process is illustrated in Figure 4.3.

The pruning can be performed in connection with any of the steps in Algorithm 4.7. However, the best place to do it is between steps 2 and 3. At this point new information from measurements has just been incorporated into the filter, and the model is just about to branch and make a deterministic time update. Hence, branching at this time is the only sensible alternative. The filter bank algorithm with pruning is given in Algorithm 4.8

**Algorithm 4.8** Filter Bank with Pruning

1. Let  $\mathcal{F}_{0|-1}$  have one single entry  $(\delta_0, x_0, \Pi_0, 1)$ , representing the initial knowledge, and  $t = 0$ .
2. Measurement update all filters in  $\mathcal{F}_{t|t-1}$  and put the result in  $\mathcal{F}_{t|t}$ . Update the mode probabilities in  $\mathcal{F}_{t|t}$  using

$$\omega_{t|t}^{(\delta_t)} = \frac{p(y_t|\delta_t)\omega_{t|t-1}^{(\delta_t)}}{\sum_{\delta_t} p(y_t|\delta_t)\omega_{t|t-1}^{(\delta_t)}} \quad (4.14a)$$

The minimum variance estimate based on this is:

$$\hat{x}_{t|t} = \sum_{\delta_t} \omega_{t|t}^{(\delta_t)} \hat{x}_{t|t}^{(\delta_t)} \quad (4.14b)$$

$$P_{t|t} = \sum_{\delta_t} \omega_{t|t}^{(\delta_t)} \left( P_{t|t}^{(\delta_t)} + (\hat{x}_{t|t}^{(\delta_t)} - \hat{x}_{t|t})(\hat{x}_{t|t}^{(\delta_t)} - \hat{x}_{t|t})^T \right). \quad (4.14c)$$

3. Prune the filter bank. Compute which mode,  $\delta$ , has at time  $t - L + 1$  the most probability stemming from it, *i.e.*,

$$\delta = \arg \max_i \sum_{\{\delta_t: \delta_{t-L+1}=i\}} \omega_{t|t}^{\delta_t}. \quad (4.14d)$$

Remove all filters from  $\mathcal{F}_{t|t}$  not having  $\delta_{t-L+1} = \delta$ , and renormalize the remaining weights,

$$\omega_{t|t}^{(i)} := \omega_{t|t}^{(i)} / \sum_j \omega_{t|t}^{(j)}. \quad (4.14e)$$

4. Branch the filter bank, *i.e.*, from  $\mathcal{F}_{t|t}$  construct  $\mathcal{F}_{t+1|t}$  by copying each entry in  $\mathcal{F}_{t|t}$   $n_\delta$  times to the new filter bank only updating the current mode, and the probability of that mode. The probability of the modes are updated using

$$\omega_{t+1|t}^{(\delta_{t+1})} = p_{\delta_{t+1}|\delta_t} \omega_{t|t}^{(\delta_t)}. \quad (4.14f)$$

The minimum variance estimate based on this is:

$$\hat{x}_{t+1|t} = \sum_{\delta_{t+1}} \omega_{t+1|t}^{(\delta_{t+1})} \hat{x}_{t+1|t}^{(\delta_{t+1})} \quad (4.14g)$$

$$P_{t+1|t} = \sum_{\delta_{t+1}} \omega_{t+1|t}^{(\delta_{t+1})} \left( P_{t+1|t}^{(\delta_{t+1})} + (\hat{x}_{t+1|t}^{(\delta_{t+1})} - \hat{x}_{t+1|t})(\hat{x}_{t+1|t}^{(\delta_{t+1})} - \hat{x}_{t+1|t})^T \right). \quad (4.14h)$$

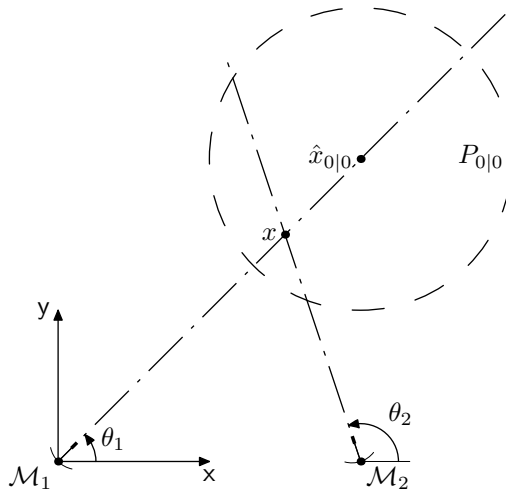
5. Time update all filters in  $\mathcal{F}_{t+1|t}$ .
6. Let  $t := t + 1$  and repeat from 2.

## 4.7 Comparison of Nonlinear Filtering Methods

In this section a bearings-only estimation problem is used to illustrate some of the properties experienced with different nonlinear extensions of the Kalman filter. First the problem setup is described and then different nonlinear filtering methods are applied.

### 4.7.1 Problem Description

The problem studied is motivated by what can be considered a somewhat strange behavior from the EKF in a situation as depicted in Figure 4.4. This is a situation where a target, known to be in an approximate location quantified by  $\hat{x}_{0|0}$  and  $P_{0|0}$ , is being triangulated using bearings-only measurements. In practice this could be an object known to be



**Figure 4.4:** Bearings-only problem. Two measurements are used, one from  $\mathcal{M}_1$  and one from  $\mathcal{M}_2$ .

stationary measured by a moving sensor. This can mathematically be described as

$$x_{t+1} = f(x_t) + w_t = x_t + w_t$$

$$y_t = h(x_t) + e_t = \arctan\left(\frac{y_t - y_t^0}{x_t - x_t^0}\right) + e_t, = \theta_t + e_t,$$

where the state  $x = (x \ y)^T$  is the Cartesian position of the target,  $w_t \equiv 0$  for clarity, and  $\text{cov}(e_t) = R$ . For this situation, the gradients needed to perform filtering using an EKF are

$$F^T := \nabla_x x = I$$

$$H^T := \nabla_x \arctan\left(\frac{y - y^0}{x - x^0}\right) = \frac{1}{(x - x^0)^2 + (y - y^0)^2} \begin{pmatrix} -(y - y^0) \\ x - x^0 \end{pmatrix}.$$

Note, the first order approximation of  $\arctan$  is best for  $|\frac{y}{x}| \gg 0$ , and that in practice moving between different quadrants is a problem that must be addressed.

Now, assume

$$x_0 = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}, \quad \hat{x}_{0|0} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad P_{0|0} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and that the bearing to the target is measured first from the position  $\mathcal{M}_1 = (0, 0)$  and then from  $\mathcal{M}_2 = (2, 0)$ , as depicted in Figure 4.4. With exact measurements of the direction to the target (however with  $\text{cov } e_t = \frac{\pi}{1000}$ , *i.e.*, the standard deviation is  $3.2^\circ$ ) the correctly inferred information about the state develops as in Figure 4.5. Especially note the slight wedge shape obtained after one measurement. The effect of this is that the probability mass moves outwards from the present estimate even though the initial estimate fits the measurement perfectly. At first this may seem somewhat counter intuitive, but it is explained by the a combination of the initial knowledge and the measurement without range information which makes it more likely that the target is farther away.

## 4.7.2 Studied Methods

The methods that will be studied with respect to their behavior in this specific setup is given below.

- The extended Kalman filter (EKF, see Section 4.4.2).
- The iterated extended Kalman filter (IEKF, see Section 4.4.3). Five iterations are used.
- The unscented Kalman filter (UKF, see Section 4.5). As suggested in [91] the following parameters are used  $\alpha = 10^{-3}$ ,  $\beta = 2$ , and  $\kappa = 0$ ,
- The particle filter (PF, see Section 4.2) with 5 000 particles.

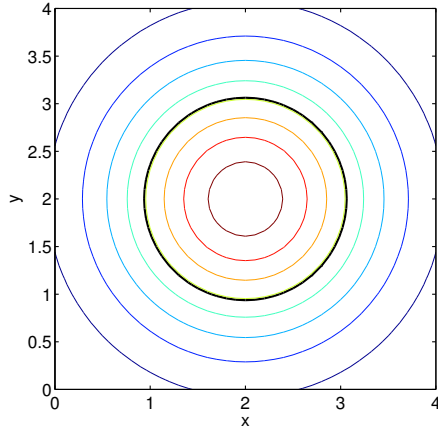
For reference the analytical Bayesian inference solution is presented in Figure 4.5.

Applying the filters yield the results in Figure 4.6. In each of the plots the estimates are shown for no measurements, one measurement from  $\mathcal{M}_1$ , and for two measurements from  $\mathcal{M}_1$  and  $\mathcal{M}_2$  in that order. Furthermore, for each estimate one standard deviation is given as a circle around the estimate. The probability of being inside one standard deviation is approximately 68% per dimension, *i.e.*, in this two dimensional setting 47%. The 47% confidence region is depicted for the analytic inference in Figure 4.7, to give a better understanding of where the most of the probability mass is located.

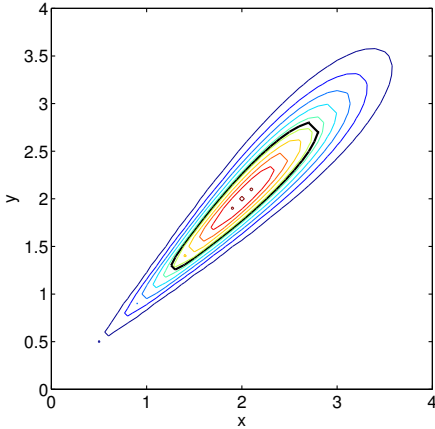
Figure 4.6 shows quite clearly the difference between the filtering methods for this specific noise realization. The performance of the EKF is not good, but the poor performance of the UKF is striking.

### No Measurement

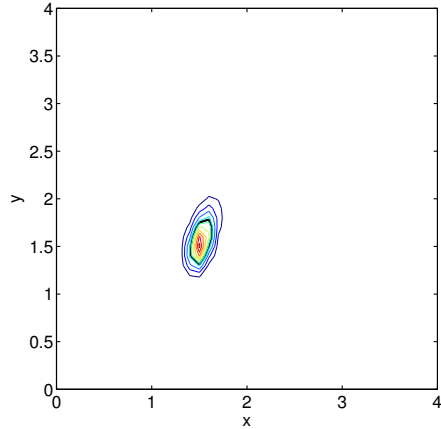
Without any information from any measurements the results from the different filters are identical — the initial knowledge is assumed to be identical.



(a) Initial knowledge, no measurements.



(b) Knowledge after one measurement from  $\mathcal{M}_1$ .

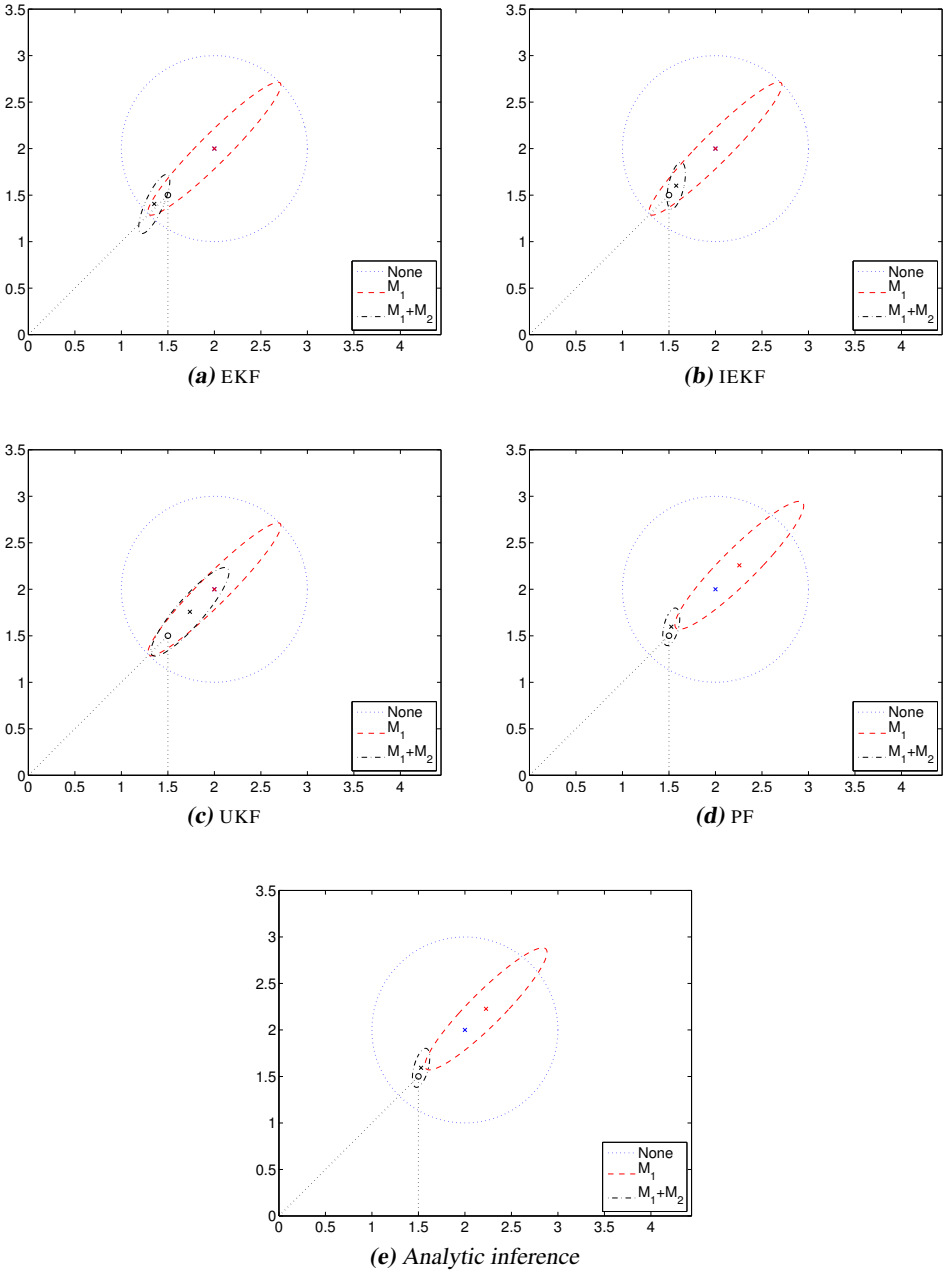


(c) Knowledge after measurements from  $\mathcal{M}_1$  and  $\mathcal{M}_2$ .

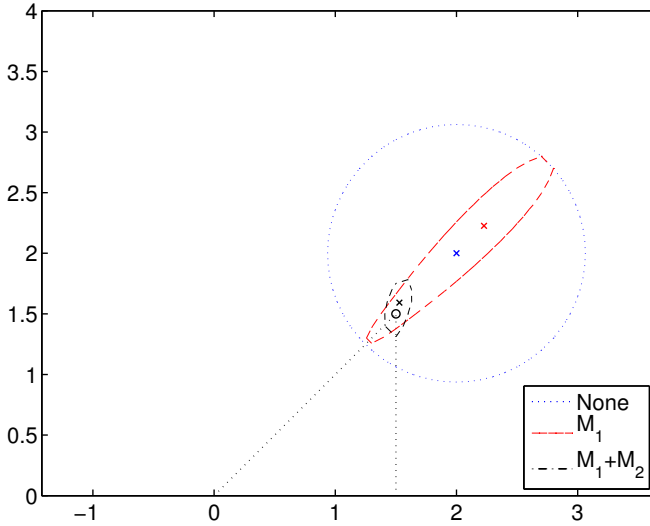
**Figure 4.5:** Analytic inferred information about  $x_0$  as given for no measurements, one measurement from  $\mathcal{M}_1 = (0, 0)$ , and one from  $\mathcal{M}_2 = (2, 0)$ . The thick lines indicate the 47% confidence regions.

**One Measurement**

With one measurement, from  $\mathcal{M}_1$ , aligned with the estimate, the EKF, the IEKF, and the UKF yield the response assumed for a linear system with Gaussian noise. The estimate remains the same and the covariance shrinks in the direction of information in the measurement. The analytic solution and the PF behaves somewhat differently, the estimate moves away from  $\mathcal{M}_1$  as more of the probability mass survives farther away from the point of the measurement. As with the other estimates the covariance matrix shrinks, and to approximately the same size. At this point the estimate given by the EKF is better



**Figure 4.6:** Estimate and covariance for the bearings-only example with no, one, and two measurements, using an EKF, an IEKF, an UKF, a PF, and true inference. (Estimates are denoted with  $\times$  in the center of each covariance ellipse, and the true target position is denoted with  $\circ$ .)



**Figure 4.7:** 47% confidence region for analytically inferred distribution. (Estimates are denoted with  $\times$  in the center of each covariance ellipse, and the true target position is denoted with  $\circ$ .)

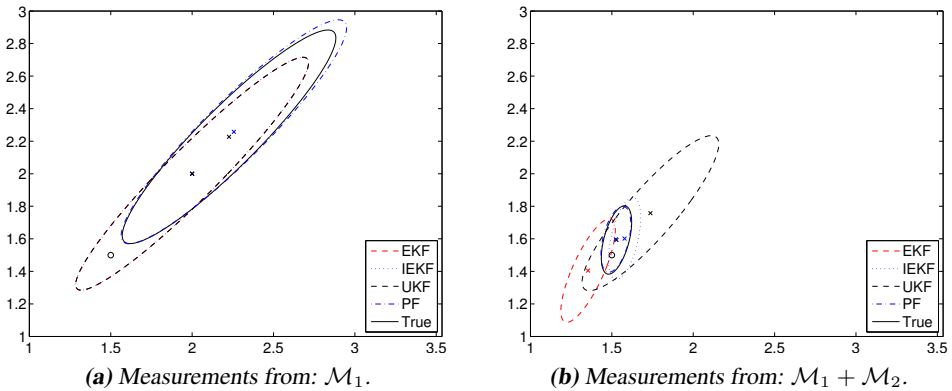
than the true interference estimate, but less coherent with the information available. The estimates from the different filters for one measurement are shown in Figure 4.8(a).

## Two Measurements

With two measurements, the first one from  $\mathcal{M}_1$  and second one from  $\mathcal{M}_2$ , the difference between the methods becomes more apparent. See Figure 4.8(b). The covariance matrix for the EKF is computed, under an incorrect linear assumption, as a linearization around the point suggested by the previous estimate. In this specific instance this means that the covariance becomes larger than expected, but also that its major axis is slightly misaligned compared to what is given by the analytic solution. Instead, the major axis of the covariance is aligned parallel to the direction from the point of measure and the expected measurement.

The idea with the IEKF is to iterate the measurement update in such a way that the best possible estimate is used for the linearization. Doing this five times pays off well, yielding an almost perfect estimate with two measurements. Both estimate and covariance is comparable to the inferred solution.

The performance of the UKF is a great disappointment. The UKF estimate is as bad as the EKF estimate, the covariance matrix is larger than with the EKF, and the misalignment is evident. The only thing that favors the UKF compared to the EKF is that the confidence region of the UKF covers the analytically inferred region, whereas the EKF covariance only covers about half of it. That is, the EKF puts too much trust in the quality of its estimate, whereas the UKF is overly pessimistic. The particle filter yields results very close to what is obtained for the true inferred state distributions. This indicates that the particle filter is



**Figure 4.8:** Comparison of covariances and estimates. (Estimates are denoted with  $\times$  in the center of each covariance ellipse, and the true target position is denoted with  $\circ$ .)

well suited for this application.

### 4.7.3 Monte Carlo Simulations

The evaluation in Section 4.7.2 suffers from just analyzing one specific noise instance. Even though the analysis gives some insight into how nonlinearities are treated by the different filters, it tells little about the actual performance of the filters, especially when it comes to choosing between the EKF, the IEKF, and the UKF.

An often used performance measure is the variance of the estimation error, often evaluated using Monte Carlo simulations. Monte Carlo simulations of the problem specified above, using the described filters, yield the result in Table 4.1(a). The table somewhat contradicts the previous results. One thing to observe is that the UKF outperforms the EKF and the IEKF. Hence, it seems that the conservative  $P$  matrix actually pays off.

Another difference compared to the theoretical analysis is that the IEKF performance is noticeable poorer than the regular EKF and the UKF. This seems to be the effect of a limited set of very bad noise instances. Looking instead at the median square error yields another result (see Table 4.1(b)). The IEKF sometimes gets stuck in updates resulting in what can be compared to positive feedback, where the new linearization actually yields a worse result than the previous one. This behavior is clearly visible if the initial estimate is moved to  $\hat{x}_0 = (1, 1)^T$ . Note that the new initial estimate is on the same bearing as the previous one as seen from the first measurement, and that the distance from the estimate to the true position is the same. Figure 4.9 depicts the estimates based on this new  $\hat{x}_0$  for the different filters. Except for the bad IEKF result nothing has changed compared to the previous result, cf. Figure 4.6. The reason for the poor IEKF performance is clearly visible in Figure 4.10. The supposedly improved estimate does actually produce a worse estimate when used for a new linearization. This estimate when used again produce an even worse estimate, and so on. In this case, the further from the point of the measurement the estimate is, the larger  $K$  becomes in the EKF. A larger  $K$  yields an estimate even

**Table 4.1:** Filter performance for 1 000 Monte Carlo simulations. Mean square error and median square error.

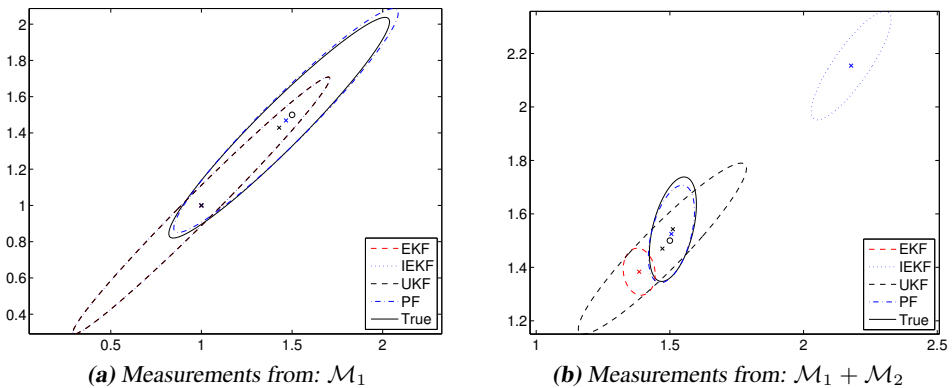
(a) Mean square error.

Filter	Measurements		
	None	$\mathcal{M}_1$	$\mathcal{M}_1 + \mathcal{M}_2$
True	2.01	0.95	0.06
EKF	2.01	1.33	1.23
IEKF	2.01	7.18	8.43
UKF	2.01	1.33	0.95
PF	2.01	0.95	0.06

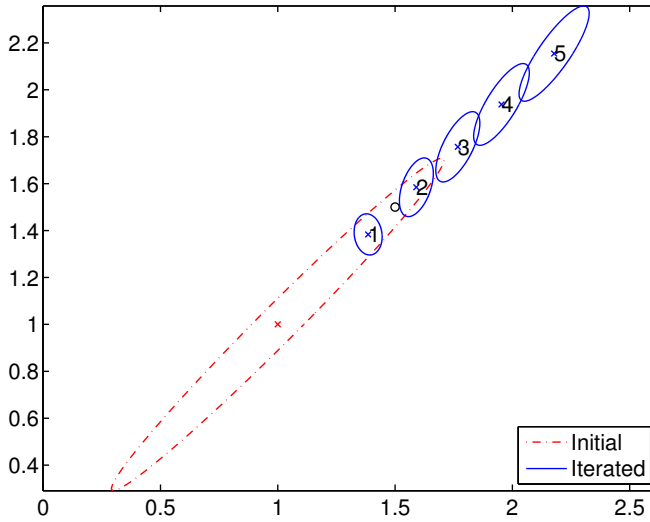
(b) Median square error.

Filter	Measurements		
	None	$\mathcal{M}_1$	$\mathcal{M}_1 + \mathcal{M}_2$
True	1.42	0.38	0.03
EKF	1.42	0.64	0.13
IEKF	1.42	0.47	0.07
UKF	1.42	0.64	0.17
PF	1.42	0.39	0.03

further away from the measurement point. This induces an even larger  $K$  and so on. This is just one example, there are other combinations that produce much worse results. It turns out that the comparable poor IEKF performance is introduced by a few of these degenerate cases.



**Figure 4.9:** Estimate with  $\hat{x}_0 = (0, 0)^T$ , based on one and two measurements. (Estimates are denoted with  $\times$  in the center of each covariance ellipse, and the true target position is denoted with  $\circ$ .)



**Figure 4.10:** Illustration of the effect of the iteration in the IEKF when the second measurement arrives. (Estimates are denoted with  $\times$  in the center of each covariance ellipse, and the true target position is denoted with  $\circ$ .)

Finally, note that the PF and the inferred distribution is almost identical. Worth noticing, though, is the substantially better estimates achieved with the PF compared to the other used filters. Hence, this is a situation when the PF pays off.

# 5

---

## Cramér-Rao Lower Bound

IT IS OFTEN of interest to know how well a parameter or a state can be estimated since that can answer questions such as: Is it possible to improve the present estimation performance with a more advanced estimator? Is it worthwhile to spend more time tuning the filter? Is it theoretically possible to fulfill these accuracy requirements? These questions are both important and common. The *Cramér-Rao lower bound* (CRLB) is a measure that can be used to answer these questions. As mentioned in Section 2.1.2, the CRLB is a lower bound on the variance of any unbiased estimator  $\hat{\theta}$  of a parameter, given by the inverse Fisher information,

$$\text{cov}(\hat{\theta}) \succeq \mathcal{I}^{-1}(\theta) = P_{\theta}^{\text{CRLB}}.$$

This chapter extends the CRLB to describe estimation of the state in dynamic systems; first for deterministic trajectories, where the *parametric Cramér-Rao lower bound* can be used, and then for stochastic trajectories using the theory for the *posterior Cramér-Rao lower bound*. It is then shown that for linear systems, the parametric and the posterior CRLB bounds yield identical limits in terms of the intrinsic accuracy of the involved noise distributions. Comparing the CRLB for the information in the true noise PDF and a Gaussian approximation makes it possible to see how much can be gained using nonlinear estimation methods. Finally, Monte Carlo simulations for two different systems exemplify the theory.

### 5.1 Parametric Cramér-Rao Lower Bound

The *parametric Cramér-Rao lower bound* is a natural extension to the CRLB for parameter estimation. Basically, treat the state at each time as a parameter with a deterministic but unknown value (more precisely treat each process noise realization,  $w_t$ , as a parameter from which the state can then be found) and derive a bound for how well these parameters

can be estimated. The subject is thoroughly treated in e.g., [8, 83], therefore only the resulting theorem is given here.

**Theorem 5.1 (Parametric Cramér-Rao lower bound)**

*The parametric Cramér-Rao lower bound for a one-step-ahead prediction*

$$P_{t+1|t} \preceq \mathbb{E}_{\hat{x}_{t+1|t}} \left( (\hat{x}_{t+1|t} - x_{t+1}^0)(\hat{x}_{t+1|t} - x_{t+1}^0)^T \right)$$

and filtering,

$$P_{t|t} \preceq \mathbb{E}_{\hat{x}_{t|t}} \left( (\hat{x}_{t|t} - x_t^0)(\hat{x}_{t|t} - x_t^0)^T \right),$$

for the system (3.2) is given by the recursion:

$$\begin{aligned} P_{t|t} &= P_{t|t-1} - P_{t|t-1} H_t^T (H_t P_{t|t-1} H_t^T + R_t)^{-1} H_t P_{t|t-1} \\ P_{t+1|t} &= F_t P_{t|t} F_t^T + G_t Q_t G_t^T, \end{aligned}$$

initialized with  $P_{0|-1}^{-1} = \mathcal{I}_{x_0}$  and with

$$\begin{aligned} F_t^T &= \nabla_{x_t} f(x_t, w_t^0) \Big|_{x_t=x_t^0}, & G_t^T &= \nabla_{w_t} f(x_t^0, w_t) \Big|_{w_t=w_t^0}, \\ H_t^T R_t^{-1} H_t &= -\mathbb{E}_{y_t} \left( \Delta_{x_t}^T p(y_t|x_t) \right), & Q_t^{-1} &= -\mathbb{E}_{x_t} \left( \Delta_{w_t}^T p(x_t|w_t^0) \right), \end{aligned}$$

where superscript  $^0$  denotes the true value, and the factorization between  $H_t$  and  $R_t$  is chosen such that  $R_t \succ 0$ .

The bounds are valid if all involved gradients, expectations, and matrix inverses exist. Explicitly this means that  $p(x_{t+1}|x_t)$  must be well defined for all  $t$ .

**Proof:** For the prediction bound, see Theorem 4.3 in [8]. The proof of the filtering bound follows with only minor modifications.  $\square$

The recursions in Theorem 5.1 should be recognized as the *Riccati* recursions used in the EKF to update the error covariance, (4.9). The difference lies not in the mathematical expressions themselves, but in the exact matrices involved. The matrices are however related, as the naming scheme suggests. If all involved distributions are Gaussian, the expressions are exactly the same as for the EKF if it is fed with the correct state trajectory. This connection is shown in [83], and is useful when it comes to evaluating a filter compared to the parametric CRLB. Simply run the EKF and feed it with the correct states to get the parametric CRLB. With the CRLB available Monte Carlo simulations is a common technique to derive the performance of the filter. The idea is to use Monte Carlo integration over enough different measurement noise instances to use the *mean square error* (MSE) instead of the variance of the error. One way to do this is given in Algorithm 5.1.

The results in Theorem 5.1 are in [8] extended to cover multi-step prediction and smoothing, too. The results are similar to those presented above in their resemblance to the covariance propagation in the EKF.

**Algorithm 5.1** Parametric Mean Square Error

1. Generate *one* true state trajectory,  $\mathbb{X}$ .
2. Based on the trajectory  $\mathbb{X}$ , generate  $M \gg 1$  independent measurement sequences,  $\mathbb{Y}^{(i)}$  for  $i = 1, \dots, M$ .
3. Apply the filter to the measurements to get  $M$  estimated trajectories,  $\hat{\mathbb{X}}^{(i)}$ .
4. How well the filter performs compared to the parametric CRLB,  $P^{\text{CRLB}}$ , is now given by,

$$\frac{1}{M} \sum_{i=1}^M (\hat{x}_t^{(i)} - x_t)(\hat{x}_t^{(i)} - x_t)^T \underset{\sim}{\succ} P_t^{\text{CRLB}},$$

where  $\underset{\sim}{\succ}$  indicates that for limited  $M$  the MSE is only approximately bounded by the CRLB.

## 5.2 Posterior Cramér-Rao Lower Bound

The parametric CRLB is insufficient in a Bayesian estimation setting since it assumes a true state trajectory, which is inconsistent with the Bayesian philosophy. The introduction of the *posterior Cramér-Rao lower bound* solves this philosophical dilemma. The posterior CRLB differs from the parametric CRLB in that it does not assume a true trajectory, hence it is acceptable from a Bayesian perspective. The posterior CRLB is more frequently treated in the literature, some references are [12, 18, 23, 24, 85]. As with the parametric CRLB, [8] provides a through description and derivation of the posterior CRLB. Therefore, Theorem 5.2 is given here without further derivation.

### Theorem 5.2 (Posterior Cramér-Rao lower bound)

*The one-step-ahead posterior Cramér-Rao lower bound for the model (3.2) is given by the recursion*

$$P_{t+1|t}^{-1} = Q_t^{-1} - S_t^T (P_{t|t-1}^{-1} + R_t^{-1} + V_t)^{-1} S_t$$

*initiated with  $P_{0|-1}^{-1} = \mathcal{I}_{x_0}^{-1}$ , and the filtering bound by*

$$P_{t+1|t+1}^{-1} = Q_t^{-1} + R_{t+1}^{-1} - S_t^T (P_{t|t}^{-1} + V_t)^{-1} S_t$$

*initiated with  $P_{0|0}^{-1} = (P_{0|-1}^{-1} + R_0^{-1})^{-1}$ , both recursions use the system dependent matrices:*

$$\begin{aligned} V_t &= -\mathbb{E}_{x_t, w_t} \left( \Delta_{x_t}^{x_t} \log p(x_{t+1}|x_t) \right), & R_t^{-1} &= -\mathbb{E}_{x_t, y_t} \left( \Delta_{x_t}^{x_t} \log p(y_t|x_t) \right), \\ S_t &= -\mathbb{E}_{x_t, w_t} \left( \Delta_{x_t}^{x_{t+1}} \log p(x_{t+1}|x_t) \right), & Q_t^{-1} &= -\mathbb{E}_{x_t, w_t} \left( \Delta_{x_{t+1}}^{x_{t+1}} \log p(x_{t+1}|x_t) \right). \end{aligned}$$

*For the bounds to be valid the involved differentiations and expectations must exist, however there is no need for  $R_t$  and  $Q_t$  to exist as long as the matrices  $R_t^{-1}$  and  $Q_t^{-1}$  do. Explicitly this means that  $p(x_{t+1}|x_t)$  must be well defined for all  $t$ .*

**Proof:** See the proof of Theorem 4.5 in [8]. □

As with Theorem 5.1, [8] elaborates further on the posterior CRLB and gives expressions for multi-step prediction and smoothing.

One observation that should be made is that the posterior CRLB is often harder to evaluate than the parametric because extra stochastic dimensions are introduced. However, in some special cases can explicit expressions be found. One such case the systems with linear dynamics. For situations when no explicit posterior CRLB expressions exist, [84] suggests a method based on Monte Carlo simulations, similar to the particle filter, to compute the CRLB numerically.

To evaluate a filter against the posterior CRLB Monte Carlo simulations can be used in a way similar to what is described for the parametric CRLB. However, since the trajectories are also random, different trajectories must be generated for each measurement sequence. See Algorithm 5.2.

---

**Algorithm 5.2** Posterior Mean Square Error

---

1. Generate  $M \gg 1$  independent trajectories,  $\mathbb{X}^{(i)}$  for  $i = 1, \dots, M$ .
2. Generate  $M$  measurement sequences, where  $\mathbb{Y}^{(i)}$  is based on the trajectory  $\mathbb{X}^{(i)}$ .
3. Apply the filter to the measurements to get  $M$  estimated trajectories,  $\hat{\mathbb{X}}^{(i)}$ .
4. How well the filter performs compared to the parametric CRLB,  $P^{\text{CRLB}}$ , is now given by,

$$\frac{1}{M} \sum_{i=1}^M (\hat{x}_t^{(i)} - x_t^{(i)})(\hat{x}_t^{(i)} - x_t^{(i)})^T \lesssim P_t^{\text{CRLB}},$$

where  $\lesssim$  indicate that for limited  $M$  the MSE is only approximately bounded by the CRLB.

---

### 5.3 Cramér-Rao Lower Bounds for Linear Systems

The sequel of this chapter is devoted to studies of the parametric and posterior CRLB of linear systems, *i.e.*, models of the type introduced in Section 3.1.2. Linear models are studied because they allow for analytical analysis and because the Kalman filter offers the BLUE for these systems. Hence, it is easy to compare optimal linear performance to the performance bound given by the CRLB. The understanding of the linear models derived this way can then be used to get better understanding of nonlinear systems, where the mathematics is usually more involved and less intuitive.

First, recall the linear model (3.3) (here without any deterministic input to simplify the notation),

$$x_{t+1} = F_t' x_t + G_t' w_t, \tag{5.1a}$$

$$y_t = H_t' x_t + e_t, \tag{5.1b}$$

with  $\text{cov}(w_t) = Q'_t$  and  $\text{cov}(e_t) = R'_t$ . Note the 's used to distinguish between the system parameters and the quantities used in the CRLB expressions.

For this system, derive an expression for the parametric CRLB using Theorem 5.1. The expressions needed are:

$$\begin{aligned} F_t^T &= \nabla_{x_t} f(x_t, w_t^0) \Big|_{x_t=x_t^0} = F_t'^T \\ G_t^T &= \nabla_{w_t} f(x_t^0, w_t) \Big|_{w_t=w_t^0} = G_t'^T \\ H_t^T R_t^{-1} H_t &= -\mathbb{E}_{y_t} \left( \Delta_{x_t}^{x_t} p(y_t|x_t) \right) = H_t'^T \mathcal{I}_{e_t} H_t' \\ Q_t^{-1} &= -\mathbb{E}_{x_t} \left( \Delta_{w_t}^{w_t} p(x_t|w_t^0) \right) = \mathcal{I}_{w_t}. \end{aligned}$$

In conclusion, for linear Gaussian systems, the naming convention in the linear model (3.3) maps directly onto the parametric CRLB notation in Theorem 5.1. For linear non-Gaussian systems, the system matrices still map directly, however, the covariances in the system model should be substituted for the inverse information of the stochastic variable. The inverse can in a way be interpreted as the effective covariance in a filter which is optimal in the sense that it reaches the CRLB.

Using the expressions to calculate the CRLB yields the recursion

$$\begin{aligned} P_{t|t} &= P_{t|t-1} - P_{t|t-1} H_t'^T (H_t' P_{t|t-1} H_t'^T + \mathcal{I}_{e_t}^{-1})^{-1} H_t' P_{t|t-1} \\ &= (P_{t|t-1}^{-1} + H_t'^T \mathcal{I}_{e_t} H_t')^{-1}, \end{aligned} \quad (5.2a)$$

$$P_{t+1|t} = F_t' P_{t|t} F_t'^T + G_t' \mathcal{I}_{w_t}^{-1} G_t'^T \quad (5.2b)$$

initiated with  $P_{0|-1} = \mathcal{I}_{x_0}^{-1}$ . The second form of  $P_{t|t}$  follows from applying the matrix inversion lemma once. Note that this is the same Riccati recursion as used for the error covariances in the Kalman filter.

Next, the posterior CRLB for the linear model (5.1) is derived. The expressions used in Theorem 5.2 simplifies to

$$\begin{aligned} Q_t^{-1} &= \mathcal{I}_{\bar{w}_t} = (G_t' \mathcal{I}_{w_t} G_t')^{-1}, \\ R_t^{-1} &= H_t'^T \mathcal{I}_{e_t} H_t', \\ S_t &= -F_t'^T \mathcal{I}_{\bar{w}_t} = -F_t'^T (G_t' \mathcal{I}_{w_t} G_t')^{-1}, \\ V_t &= F_t'^T \mathcal{I}_{w_t} F_t' = F_t'^T (G_t' \mathcal{I}_{w_t} G_t')^{-1} F_t', \end{aligned}$$

where  $\bar{w}_t := G_t w_t$ . Detailed derivations, in terms of  $\mathcal{I}_{\bar{w}_t}$ , can be found in [8], and the last equalities follows from applying Theorem 2.2. Substituting these into the posterior CRLB expressions yields,

$$\begin{aligned} P_{t+1|t}^{-1} &= \mathcal{I}_{\bar{w}_t} - (-F_t'^T \mathcal{I}_{\bar{w}_t})^T (P_{t|t-1}^{-1} + H_t'^T \mathcal{I}_{e_t} H_t' + F_t'^T \mathcal{I}_{\bar{w}_t} F_t')^{-1} (-F_t'^T \mathcal{I}_{\bar{w}_t}) \\ &= (\mathcal{I}_{\bar{w}_t}^{-1} + F_t' (P_{t|t-1}^{-1} + H_t'^T \mathcal{I}_{e_t} H_t')^{-1} F_t'^T)^{-1} \\ &= (G_t \mathcal{I}_{w_t}^{-1} G_t^T + F_t' (P_{t|t-1}^{-1} + H_t'^T \mathcal{I}_{e_t} H_t')^{-1} F_t'^T)^{-1} \end{aligned} \quad (5.3a)$$

for the prediction bound and for the filtration bound

$$\begin{aligned}
 P_{t|t}^{-1} &= \mathcal{I}_{\bar{w}_t} + H_t'^T \mathcal{I}_{e_t} H_t' - (-F_t'^T \mathcal{I}_{\bar{w}_t})^T (P_{t-1|t-1}^{-1} + F_t'^T \mathcal{I}_{\bar{w}_t} F_t')^{-1} (-F_t'^T \mathcal{I}_{\bar{w}_t}) \\
 &= H_t'^T \mathcal{I}_{e_t} H_t' + (\mathcal{I}_{\bar{w}_t}^{-1} + F_t' P_{t-1|t-1} F_t'^T)^{-1} \\
 &= H_t'^T \mathcal{I}_{e_t} H_t' + (G_t \mathcal{I}_{w_t}^{-1} G_t^T + F_t' P_{t-1|t-1} F_t'^T)^{-1}. \tag{5.3b}
 \end{aligned}$$

Inspection of the expressions reveals that the same recursive expression as for the parametric CRLB can be used to calculate the posterior CRLB for linear systems. Hence, for linear systems do the parametric CRLB and posterior CRLB yield exactly the same bounds. That is, in this case the results obtained using a classical view of estimation and a Bayesian view are the same. The result follows because the state does not change the dynamics or the measurement equation of the system, hence not knowing the state will not affect the estimation. If the state influences the dynamics or measurements of the system then the uncertainty of the exact system makes the estimation more difficult using the Bayesian approach. Note that for nonlinear systems, where the effective system depends on the state, the bounds should be expected to differ.

### Theorem 5.3 (Cramér-Rao Lower Bounds For Linear Systems)

For linear systems (3.3) the parametric and posterior Cramér-Rao lower bound yield the same lower bound. This lower bound is given by the recursion

$$\begin{aligned}
 P_{t|t} &= (P_{t|t-1}^{-1} + H_t^T \mathcal{I}_{e_t} H_t)^{-1}, \\
 P_{t+1|t} &= F_t P_{t|t} F_t^T + G_t \mathcal{I}_{w_t}^{-1} G_t^T,
 \end{aligned}$$

initiated with  $P_{0|-1} = \mathcal{I}_{x_0}^{-1}$ .

**Proof:** The result follows from Theorems 5.1 and 5.2 and the derivation of the parametric and the posterior CRLB for a linear system as done above.  $\square$

### Corollary 5.1

If the linear system is time invariant and there exist a bounded asymptotic Cramér-Rao lower bound then the prediction bound  $\bar{P}_+$  is

$$\bar{P}_+ = F(\bar{P}_+^{-1} + H_t^T \mathcal{I}_{e_t} H_t)^{-1} F^T + G \mathcal{I}_{w_t}^{-1} G^T,$$

and the filtering bound  $\bar{P}$  is

$$\bar{P} = (\bar{P}_+^{-1} + H_t^T \mathcal{I}_{e_t} H_t)^{-1}.$$

**Proof:** If asymptotic solutions exist as  $t \rightarrow \infty$ , the bounds are  $P_{t+1|t} = P_{t|t-1} =: \bar{P}_+$  and  $P_{t|t} = P_{t-1|t-1} =: \bar{P}$ . The result then follows from using  $\bar{P}_+$  and  $\bar{P}$  in Theorem 5.3.  $\square$

Conditions for when a stationary solution exists can be found in [42].

## Linear Cramér-Rao Lower Bound Properties

For linear systems the noise characteristics have a direct impact on the CRLB that is obtained. It is by studying how the changes in the intrinsic accuracy affects optimal estimation performance possible to minimize the calculations necessary to find the CRLB and to get a better understanding of the mechanisms in play.

### Lemma 5.1

For matrices  $Q_t \succeq 0$ ,  $R_t \succ 0$ , and  $P \succ 0$ , and the scalar  $\gamma > 0$  the following is true:

(i) For the time update of the Kalman filter covariance

$$\gamma(F_t P F_t^T + G_t Q G_t^T) = F_t(\gamma P)F_t^T + G_t(\gamma Q)G_t^T.$$

(ii) For the measurement update of the Kalman filter covariance

$$(P^{-1} + H_t^T R_t^{-1} H_t)^{-1} = ((\gamma P)^{-1} + H_t^T (\gamma R_t)^{-1} H_t)^{-1}.$$

(iii) If  $\tilde{Q} \preceq Q$  and  $\tilde{P} \preceq P$  this is preserved through a time update,

$$F \tilde{P} F^T + G \tilde{Q} G^T \preceq F P F^T + G Q G^T,$$

with equality if and only if  $G \tilde{Q} G^T = G Q G^T$  and  $F \tilde{P} F^T = F P F^T$ .

(iv) If  $\tilde{R} \preceq R$  and  $\tilde{P} \preceq P$  the relation is preserved through a measurement update,

$$(\tilde{P}^{-1} + H^T \tilde{R}^{-1} H)^{-1} \preceq (P^{-1} + H^T R^{-1} H)^{-1},$$

with equality if and only if  $H^T \tilde{R}^{-1} H = H^T R^{-1} H$  and  $\tilde{P} = P$ .

**Proof:** The properties (i)–(ii) can be shown by factoring out  $\gamma$  in right-hand side expressions.

To show property (iii) compare the two expressions,

$$(F P F^T + G Q G^T) - (F \tilde{P} F^T + G \tilde{Q} G^T) = F(P - \tilde{P})F^T + G(Q - \tilde{Q})G^T \succeq 0,$$

since  $\tilde{P} \preceq P$  and  $\tilde{Q} \preceq Q$ .

Property (iv) follows in a similar manner,

$$\begin{aligned} (P^{-1} + H^T R^{-1} H) - (\tilde{P}^{-1} + H^T \tilde{R}^{-1} H) \\ = (P^{-1} - \tilde{P}^{-1}) + H^T (R^{-1} - \tilde{R}^{-1}) H \succeq 0, \end{aligned}$$

since  $A \succ 0$  and  $B \succ 0$  implies  $A \succeq B \Leftrightarrow A^{-1} \preceq B^{-1}$ , and  $P \succeq \tilde{P}$  and  $Q \succeq \tilde{Q}$ . This in turn provides the inequality.  $\square$

Lemma 5.1 can be used to show what happens with the CRLB for linear systems when the information in the noise changes; properties (i) and (iii) describe how the state estimate loses information in the time update step if a less informative process noise is

introduces, and properties (ii) and (iv) how information is lost when introducing less informative measurements. Common for both time and measurement updates is that if all involved stochastic contributions are scaled then the result of the update is scaled with the same factor. Another result is that if the information content in a stochastic variable is increased the CRLB for the estimate is improved as well, unless the information is gained in a direction that is neglected by the system. However, introducing more informative noise will never have a negative effect on the CRLB. Furthermore, if a system is affected by non-Gaussian noise the CRLB is likely to improve since according to Theorem 2.1 the Gaussian distribution is the least informative distribution.

The results above deals with individual steps in recursive formulas used to calculate the Kalman filter and CRLB covariance matrices. By repeatedly applying Lemma 5.1 results for the CRLB at a specific time can be derived.

#### Theorem 5.4

Assume a linear system (3.3), with inverse intrinsic noise accuracies  $Q_\tau \succeq 0$ ,  $R_\tau \succ 0$ , and  $P_{0|-1} \succeq 0$  for  $\tau \leq t$ , and for which the Cramér-Rao lower bounds at time  $t$  are  $P_{t|t}$  and  $P_{t+1|t}$ . The following two properties are true.

- If all  $Q_\tau$  and  $R_\tau$  are changed to  $\gamma Q_\tau$  and  $\gamma R_\tau$ , respectively, and  $P_{0|-1}$  to  $\gamma P_{0|-1}$ , the CRLB of the resulting system are  $\gamma P_{t|t}$  and  $\gamma P_{t+1|t}$ .
- For a new system with the same dynamics as the first and with  $Q_\tau \succeq \tilde{Q}_\tau$ ,  $R_\tau \succeq \tilde{R}_\tau$ , and  $P_{0|-1} \succeq \tilde{P}_{0|-1}$  the CRLB of the two systems relates as  $P_{t|t} \succeq \tilde{P}_{t|t}$  and  $P_{t+1|t} \succeq \tilde{P}_{t+1|t}$  where the  $\tilde{\cdot}$  indicate a property of the new system.

**Proof:** Recursively apply Lemma 5.1. □

The first part of Theorem 5.4 states that if all intrinsic accuracies in the system is scaled with the same factor, i.e., have the same relative accuracy, the resulting CRLB will also be scaled with the same factor. The second part states that if any intrinsic accuracy is improved, i.e., a noise becomes more informative, this can only improve the resulting CRLB and unless the direction of the improvement is ignored by the system the CRLB will improve. Since non-Gaussian noise is more informative than Gaussian noise, having non-Gaussian distributions in the system description will likely improve the estimation performance.

For time invariant linear systems the stationary performance is given by the Riccati equation if a stationary bound exists.

#### Corollary 5.2

If the linear system is time invariant and an asymptotic Cramér-Rao lower bound exists, denote  $\bar{\kappa}_+(Q, R) := P_{t+1|t}$  and  $\bar{\kappa}(Q, R) := P_{t|t}$  as  $t \rightarrow +\infty$ , where  $Q$  and  $R$  are the parameters in the CRLB recursion, then:

- For  $\gamma > 0$

$$\gamma \bar{\kappa}_+(Q, R) = \bar{\kappa}_+(\gamma Q, \gamma R) \quad \text{and} \quad \gamma \bar{\kappa}(Q, R) = \bar{\kappa}(\gamma Q, \gamma R).$$

- If  $\tilde{Q} \preceq Q$  and  $\tilde{R} \preceq R$  then

$$\bar{\kappa}_+(\tilde{Q}, \tilde{R}) \preceq \bar{\kappa}_+(Q, R) \quad \text{and} \quad \bar{\kappa}(\tilde{Q}, \tilde{R}) \preceq \bar{\kappa}(Q, R).$$

**Proof:** Follows immediately from Theorem 5.4 given that asymptotic solutions exist.  $\square$

Using the notation in Corollary 5.2,  $\bar{\kappa}(Q, R)$  denotes the asymptotic CRLB for a (implicitly defined) linear time-invariant system if  $Q$  and  $R$  are the inverse intrinsic accuracies of the process and measurement noise, respectively, and the BLUE performance if  $Q$  and  $R$  are the covariances. At the same time is  $\bar{\kappa}(Q, R)$  the solution to a Riccati which makes it fairly easy to compute with standard methods. This makes  $\bar{\kappa}(Q, R)$  a good quantity to study when deciding if a nonlinear filtering method should be evaluated or not.

## 5.4 Applications of the Theory

The remaining part of this chapter will be used to illustrate the Cramér-Rao lower bound theory with simulations. Two examples will be used, the first is a *constant velocity* (CV) model, and the second is a model of a DC motor. The effect of introducing non-Gaussian noise as process noise and measurement noise is studied.

### 5.4.1 Constant Velocity Model

The *constant velocity* (CV) model, previously described in Example 3.1,

$$x_{t+1} = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix} x_t + \begin{pmatrix} \frac{1}{2}T^2 \\ T \end{pmatrix} w_t, \quad (5.4)$$

$$y_t = (1 \quad 0) x_t + e_t, \quad (5.5)$$

with  $w_t$  and  $e_t$  mutually independent white noises. Here with  $Q = R = 1$  and  $T = 1$ . The state is constructed as

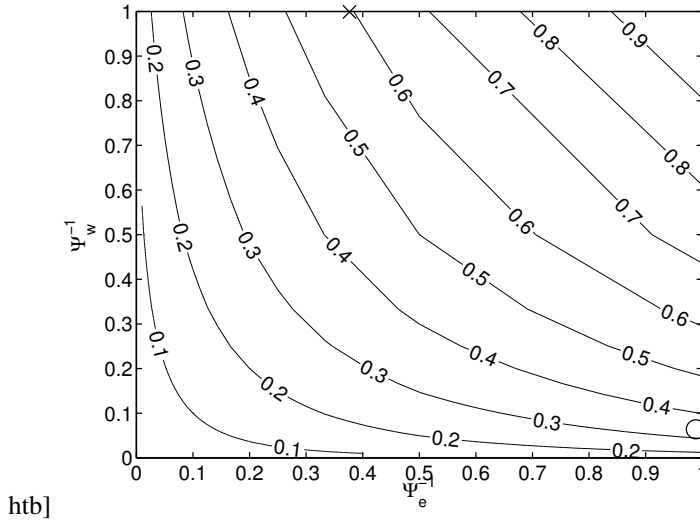
$$x = \begin{pmatrix} x \\ v \end{pmatrix},$$

where  $x$  and  $v$  represent position and velocity, respectively. This is, however, just one possible interpretation of the state-space model, it can also be interpreted as a second order random walk or a double integration and can then be used to model a slowly changing state, e.g., a fluid level in a container.

With this model and with the given noise levels, the asymptotic variance of the position component,  $x$ , of the state vector of a BLUE prediction is  $\text{var}(\hat{x}_{t+1|t}) = \bar{\kappa}_+^x(1, 1) = 3.0$ , where  $\bar{\kappa}_+(\cdot, \cdot)$  is defined in Corollary 5.2. However, if the noise is non-Gaussian the Cramér-Rao lower bound is lower. Since both  $w_t$  and  $e_t$  are scalar it is possible to illustrate how the optimal performance, normalized with  $\bar{\kappa}_+^x(1, 1)$ , varies with the intrinsic accuracy of the noise. In this case is the intrinsic accuracy equal to the relative accuracy due to the unit variance for all noise terms. The prediction performance

$$\frac{\bar{\kappa}_+^x(\mathcal{I}_{w_t}^{-1}, \mathcal{I}_{e_t}^{-1})}{\bar{\kappa}_+^x(\text{var}(w_t), \text{var}(e_t))} = \frac{\bar{\kappa}_+^x(\Psi_{w_t}^{-1}, \Psi_{e_t}^{-1})}{\bar{\kappa}_+^x(1, 1)}$$

is in Figure 5.1 presented as a function of the relative accuracies involved. Recall the definition of relative accuracy in Section 2.1.2 as a measure of how much more useful information is contained in a distribution compared to a Gaussian distribution. Hence, Figure 5.1 is in much similar to the relative accuracies in the Figures 2.3(a), 2.4(a), and 2.5(a)



**Figure 5.1:** Contour plot of the optimal filter performance, as a function of the relative accuracies  $\Psi_{w_t}$  and  $\Psi_{e_t}$ ,  $\bar{\kappa}_+^x(\Psi_{w_t}^{-1}, \Psi_{e_t}^{-1})/\bar{\kappa}_+^x(1, 1)$  with  $\bar{\kappa}_+^x(1, 1) = 3.0$ . ( $\times$  denotes the noise in the first simulation and  $\circ$  the noise in the second simulation.)

in that they show the optimal prediction performance. From the information in the contour plot it is possible to determine if it is worthwhile to design a nonlinear filter, e.g., a particle filter, in an attempt to come closer to the CRLB. Or if the potential gain is minor and not worth the extra effort to design and implement anything but a Kalman filter, which is the BLUE. In accordance with the theoretical results above the performance improves as the relative accuracy increases, and therefore the optimal performance is acquired for maximum relative accuracy in both process noise and measurement noise. (Note the inverse on the axis in Figure 5.1, the optimal performance is obtained for  $\Psi_w^{-1} = \Psi_e^{-1} = 0$ , i.e., infinite information.) Also note the linear behavior as the relative accuracy for the process noise and measurement noise are the same. In general, when the CRLB is much lower than the BLUE performance, indicated with a low value in the plot, an nonlinear filter could pay off. However, observe that it is impossible to conclude from these results how to obtain better performance, and that the Cramér-Rao lower bound is an asymptotic result without information about how difficult it is to reach the bound, or if it is possible at all.

### Bimodal Measurement Noise

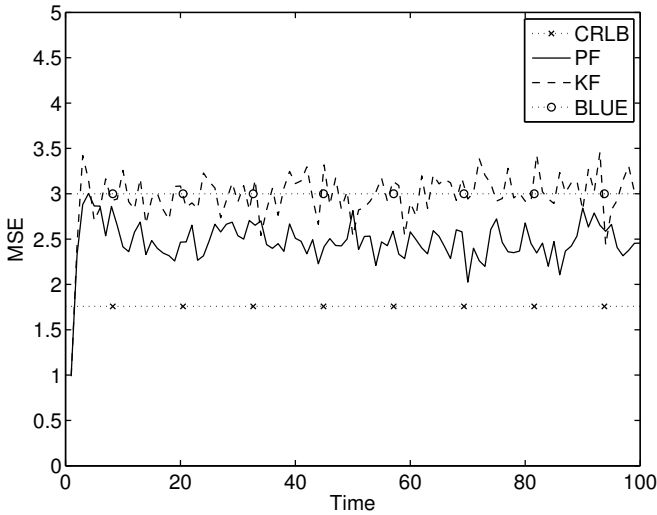
The first simulation features Gaussian process noise,  $\Psi_{w_t} = 1$ , and the non-Gaussian measurement noise,

$$e_t \sim \frac{9}{10}\mathcal{N}(0.2, 0.3) + \frac{1}{10}\mathcal{N}(-1.8, 3.7),$$

an instance of the bi-Gaussian noise presented in Section 2.3.2. The relative accuracy of the measurement noise is  $\Psi_{e_t} = 2.7$ , e.g., found in Figure 2.4(a). From Figure 5.1 (the

position denoted with  $\times$ ), or by solving the appropriate Riccati equation in Theorem 5.3, the CRLB for this system with this measurement noise can be found to be  $\bar{\kappa}_+^x(1, 0.37) = 1.8$ . That is, the optimal variance is 60% of what is obtainable with a BLUE,  $\bar{\kappa}_+^x(1, 1) = 3.0$ . Hence, the system seems to be a good candidate for a nonlinear filter.

The system is analyzed in a simulation study, where a Kalman filter and a particle filter (SIR with 50 000 particles<sup>1</sup>) are applied. The mean square error of these estimates are then computed for 1 000 Monte Carlo simulations. The MSE together with theoretical stationary limits are plotted in Figure 5.2, which shows a significant improvement when



**Figure 5.2:** MSE of 1 000 MC simulations with KF and PF (50 000 particles) on the system with bi-Gaussian measurement noise. CRLB and BLUE (variance) indicate asymptotic limits.

using the PF (approximately 18% lower variance). However, the CRLB is not reached. A reason is that the CRLB expression is asymptotic in the measurements, which leaves no guarantee that it could be reached in practice. More measurement information compared to process noise would probably improve the results, *i.e.*, more and better measurements. In practice, this could be achieved with *e.g.*, more frequent measurements and/or better sensors.

### Tri-Gaussian Process Noise

In the second simulation study with the constant velocity model the measurements are kept Gaussian, whereas the system is driven by trimodal noise, as described in Section 2.3.2,

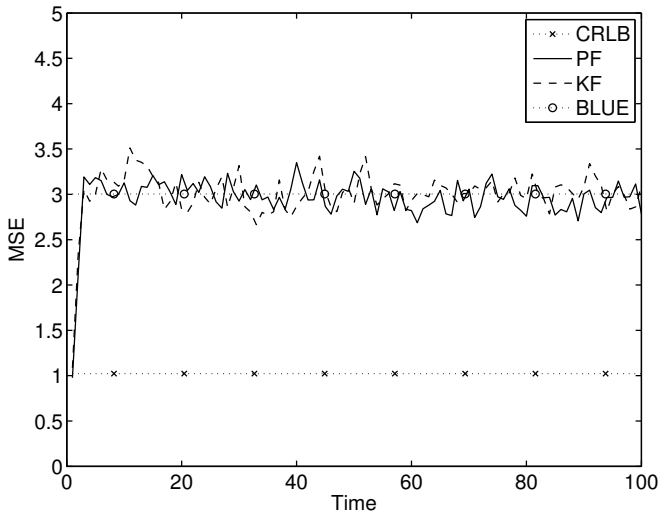
$$w_t \sim 0.075\mathcal{N}(-2.5, 0.065) + 0.85\mathcal{N}(0, 0.065) + 0.075\mathcal{N}(+2.5, 0.065),$$

<sup>1</sup>The number of particles in the PF is large, intentionally, to get the most from the filter without having to worry about numerical issues. Fewer particles could be sufficient, but this has not been analyzed.

yielding  $\Psi_e = 1$  and  $\Psi_w = 15.4$ . In this setting the mode of the noise can be interpreted as being an unknown control input with stochastic properties. The non-central modes would then represent different input to the system. As such the noise can be used in tracking application where the probability of a turn is given beforehand, but not the exact times when the target turns is unknown.

The performance of the BLUE is the same as before,  $\bar{\kappa}_x(1, 1) = 3.0$ . However, the CRLB for the system is different,  $\bar{\kappa}_x(0.065, 1) = 1.77$ . This can be derived from Figure 5.1 or by using Theorem 5.3.

Simulating this system and applying a Kalman filter and a particle filter (SIR with 50 000 particles) yields for 1 000 Monte Carlo simulations the result in Figure 5.3. In this



**Figure 5.3:** MSE for  $x$  of 1 000 MC simulations with KF and PF (50 000 particles) on the system with tri-Gaussian process noise. CRLB and BLUE (variance) indicate asymptotic limits.

simulation study it is hard to tell the difference between the particle filter and the Kalman filter, there is no significant difference. (Taking the mean over time shows that the particle filter is approximately 3% better than the Kalman filter.) The same argumentation as for the non-Gaussian measurement noise apply.

This case is also complicated by the fact that the non-Gaussian noise is only measured indirectly. As a result it is harder to extract all available information, especially for prediction since there is no measurement available to give information about the latest process noise affecting the system.

## 5.4.2 DC Motor

In this section, a DC motor (direct current motor) is used to further exemplify the Cramér-Rao lower bound theory. The DC motor is a common component in many everyday consumer products, e.g., CD-players, hard drives in computers and MP3-players, where the

accuracy of the DC motor is important. To use a DC motor for accurate positioning it is important to be able to tell the current state of the motor, *i.e.*, how much the drive shaft is turned and how fast it is turning right now. For this estimation techniques can be used. The task in this simulation study is therefore to estimate the angle and speed of the drive shaft when either the process noise, the uncertainty in the control signal, or the measurement noise is non-Gaussian. Without loss of generality the nominal input can be removed from the system, hence motivating the lack of a regular input to the system. A DC motor can be approximated with the following linear time-invariant model:

$$x_{t+1} = \begin{pmatrix} 1 & 1 - e^{-1} \\ e^{-1} & 0 \end{pmatrix} x_t + \begin{pmatrix} e^{-1} \\ 1 - e^{-1} \end{pmatrix} w_t \quad (5.6a)$$

$$y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} x_t + e_t, \quad (5.6b)$$

where  $w_t$  and  $e_t$  are independent, but not necessarily Gaussian noises.

To get a understanding of how the system is affected by non-Gaussian noise study a nominal DC motor system with the noise profile

$$\begin{aligned} w_t &\sim \mathcal{N}\left(0, \left(\frac{\pi}{180}\right)^2\right) \\ e_t &\sim \mathcal{N}\left(0, \left(\frac{\pi}{180}\right)^2\right), \end{aligned}$$

*i.e.*, both process noise and measurement noise are Gaussian and have the same order of magnitude, equivalent to a standard deviation of  $1^\circ$ . This gives the asymptotic estimation performance (in this case trace of the covariance matrix)  $\bar{\kappa}^{\text{tr}}\left(\left(\frac{\pi}{180}\right)^2, \left(\frac{\pi}{180}\right)^2\right) = 4.5 \cdot 10^{-4} \approx (1.2^\circ)^2$ . Figure 5.4 illustrates how the estimation performance is affected by non-Gaussian noise. The result is similar to what was found for the constant velocity model.

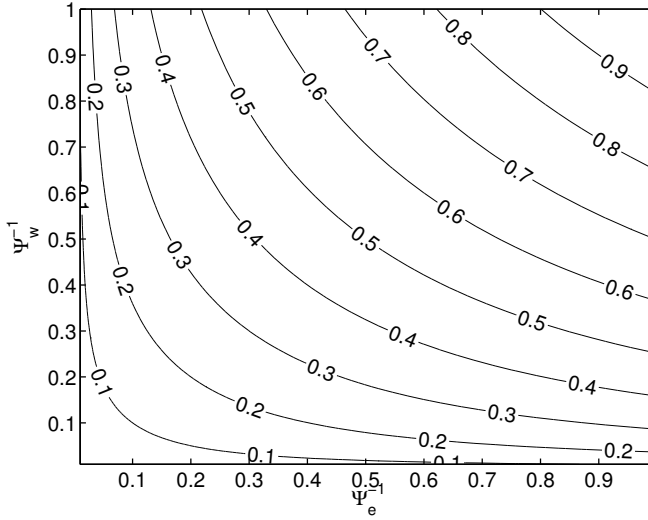
### Measurements with Outliers

For the first simulation study of the DC motor, assume the noise configuration

$$\begin{aligned} w_t &\sim \mathcal{N}\left(0, \left(\frac{\pi}{180}\right)^2\right) \\ e_t &\sim \mathcal{N}_2\left(\left(0.9, 0, \left(\frac{\pi}{180}\right)^2\right), \left(0.1, 0, \left(\frac{10\pi}{180}\right)^2\right)\right), \end{aligned}$$

where  $\mathcal{I}_e = 1.1 \cdot 10^4$  and  $\Psi_e = 9$ , and the standard deviation on each mode is  $1^\circ$ . Note that this does not give the same noise variance as was used for the nominal system above. The difference between the previously described nominal DC motor is that the measurement noise is now affected by outliers in 10% of the measurements, and the outliers have ten times the variance of the nominal measurements. In a real sensor this could be the affect of not having time enough to process all data and hence once in a while produce a measurement of low quality. The measurement noise is an instance of the noise discussed in Section 2.3.2 with scaled variance.

The high intrinsic accuracy of the measurement noise, is an indication that, unless the model is such that informative measurements does not pay off, nonlinear filtering should not be ruled out without further analysis. Computing the relative filtering performance,



**Figure 5.4:** Contour plot of the optimal filter performance, as a function of the relative accuracies  $\Psi_{w_t}$  and  $\Psi_{e_t}$ ,  $\bar{\kappa}^{\text{tr}}(\Psi_{w_t}^{-1}, \Psi_{e_t}^{-1}) / \bar{\kappa}^{\text{tr}}(\text{var}(w_t), \text{var}(e_t))$  with  $\bar{\kappa}^{\text{tr}}(\text{var}(w_t), \text{var}(e_t)) = 4.5 \cdot 10^{-4}$ .

this time the trace of the covariance matrix, for the system yields the impressive

$$\frac{\bar{\kappa}^{\text{tr}}(\mathcal{I}_w^{-1}, \mathcal{I}_e^{-1})}{\bar{\kappa}^{\text{tr}}(\text{var}(w), \text{var}(e))} = 0.5.$$

The potential performance gain is large, 50% improvement, and nonlinear filtering should definitely be evaluated for this system.

Figure 5.5 shows the result of a Monte Carlo simulation performed on this model using a Kalman filter and a particle filter (SIR with 10 000 particles<sup>2</sup>). The result is promising because the variance of the estimate comes close to the CRLB. There is still a gap between the CRLB and the performance obtained using the particle filter but it is comparably small.

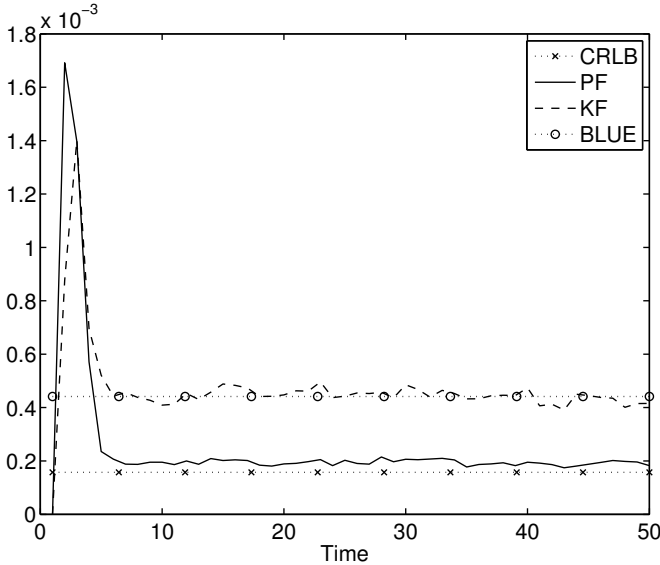
## Load Disturbances

The second DC motor system studied has the noise configuration:

$$\begin{aligned} w_t &\sim \mathcal{N}_2\left(\left(0.8, 0, \left(\frac{\pi}{180}\right)^2\right), \left(0.2, \frac{-10\pi}{180}, \left(\frac{\pi}{180}\right)^2\right)\right) \\ e_t &\sim \mathcal{N}\left(0, \left(\frac{\pi}{180}\right)^2\right). \end{aligned}$$

The bimodal process noise can be interpreted as a reoccurring load disturbance which is in affect 20% of the time resulting in a mean of the distribution at  $-10^\circ$ , e.g., it can be

<sup>2</sup>For this system 10000 particles seems to be enough to obtain maximum performance from the particle filter, increasing the number of particles to 50 000 does not improve the performance.



**Figure 5.5:** MSE for 1 000 MC simulations with KF and PF (10 000 particles) on the DC motor with outliers in the measurements. CRLB and BLUE (variance) indicate asymptotic limits.

that the drive shaft gets stuck every once in a while. The noise is a scaled instance of to the bimodal noise in Section 2.3.2. The new process noise has larger variance than the nominal system yielding  $\bar{\kappa}^{\text{tr}}(\text{var}(w_t), \text{var}(e_t)) = 3.9 \cdot 10^{-3}$ . The process noise is characterized by  $\mathcal{I}_w = 3.3 \cdot 10^3$  and  $\Psi_w = 17$ , which gives

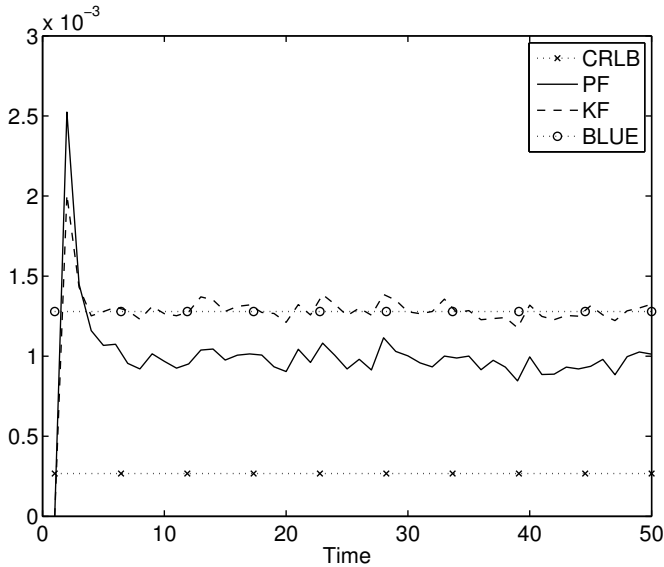
$$\frac{\bar{\kappa}^{\text{tr}}(\mathcal{I}_w^{-1}, \mathcal{I}_e^{-1})}{\bar{\kappa}^{\text{tr}}(\text{var}(w), \text{var}(e))} = 0.1.$$

Monte Carlo simulations on the system yields the result in Figure 5.6. The result is not as promising as in the outliers example since the CRLB is not reached, but there is a clear improvement in performance compared to the Kalman filter. Hence, this shows that performance can be gained with non-Gaussian process noise, even though this was not observed in the simulations of the constant velocity model with non-Gaussian process noise.

The fact that the particle filter does not reach the CRLB every time could at first come as a surprise, however the CRLB is just a lower bound without guarantees of being reachable. Lack of information in the measurements is one probable reason for this. The CRLB results are derived for asymptotic measurement information, which is not obtained for the standard filtering formulations.

### 5.4.3 Observations

The obtained performance of the four simulation studies can be concluded with the following items:



**Figure 5.6:** MSE for 1 000 MC simulations with KF and PF (10 000 particles) on the DC motor with bimodal process noise. CRLB and BLUE (variance) indicate asymptotic limits.

- It has been shown that using a nonlinear filter, in this case a particle filter, it is possible to obtain better estimation performance than with a Kalman filter if non-Gaussian noise is present.
- Computing the CRLB for the system and then comparing it to the performance of the BLUE gives an indication of how much may be gained with nonlinear filtering. However, there is no guarantee for improved performance even if the CRLB is much better than the BLUE performance.
- For the two studied systems, it has been easier to turn information from non-Gaussian measurement noise into better estimates, than it has been to utilize non-Gaussian process noise. The reasons for this have not yet been established; it may be a property of the systems studied, or a more general property as a result from how process and measurement noise enter a system.

# 6

---

## Change Detection

**D**ETEECTING AN UNEXPECTED change in a system as fast as possible is often vital in order to be able to adapt to new conditions. The change may be introduced by a component failure, as an effect of constant usage, or simply be the result of a change in the surrounding environment. Trying to detect a change in a system is called *change detection* or *fault detection* since a change in a system is often considered to be a fault. The main objective in this chapter is to determine if a change has occurred. This chapter can be read without first having read Chapters 4 and 5, however the material in those chapters increases the level of understanding of the detection theory.

In this chapter,  $f_t$  denotes the deterministic but unknown parameter that may change. The models used here were presented in Section 3.1.3, where

$$x_{t+1} = f(x_t, w_t, u_t, f_t) \tag{6.1a}$$

$$y_t = h(x_t, u_t, f_t) + e_t, \tag{6.1b}$$

is the most general model treated. Hypothesis testing is introduced first in this chapter since it, in one way or another, is the base for all fault detection. Once the basic concepts are known, general techniques for detection and optimality results are introduced. Results for linear system are then derived and the theory is illustrated using Monte Carlo simulations.

### 6.1 Hypothesis Testing

The foundation for change detection is *hypothesis testing*. Hypothesis testing is treated in depth by many text books on change detection, e.g., [5, 48, 59, 89] just to mention a few, and is used to decide between statistical statements called *hypotheses*. Hypotheses are in this thesis denoted with  $\mathcal{H}_i$ , where  $\mathcal{H}_0$  is the *null hypothesis* which is put against one or several *alternate hypotheses*.

**Definition 6.1.** A hypothesis,  $\mathcal{H}$ , is any assumption about a distribution. The hypothesis  $\mathcal{H}$  is called a *simple hypothesis* if it reduces to a single point in the parameterization of the distribution, and otherwise  $\mathcal{H}$  constitutes a *composite hypothesis*.

---

**Example 6.1: Hypothesis**

---

Assume that  $y$  is a measurement of a parameter  $f$  in presence of Gaussian noise, i.e.,  $y \sim \mathcal{N}(f, 1)$ , and that either  $f = 0$  or  $f = 1$ . Deciding which value  $f$  has is to decide between two simple hypotheses:

$$\begin{cases} \mathcal{H}_0 : & \text{If } y \sim \mathcal{N}(0, 1), \text{ i.e., } f = 0, \\ \mathcal{H}_1 : & \text{If } y \sim \mathcal{N}(1, 1), \text{ i.e., } f = 1. \end{cases}$$

Now assume instead that the question is whether  $f = 0$  or not, and that  $f$  can take on any real value. The two new hypotheses to decide between are then

$$\begin{cases} \mathcal{H}_0 : & \text{If } y \sim \mathcal{N}(0, 1), \text{ i.e., } f = 0, \\ \mathcal{H}_1 : & \text{If } y \sim \mathcal{N}(f, 1), \text{ i.e., } f \neq 0, \end{cases}$$

where the alternative hypothesis,  $\mathcal{H}_1$ , is a composite hypothesis since it corresponds to several different values of  $f$ , and the null hypothesis,  $\mathcal{H}_0$ , is simple. As will be seen further on, the second set of hypotheses is much harder to handle than the first one due to the less informative composite hypothesis.

---

The general principle for deciding between the hypotheses is to derive a test statistic,  $L(\mathbb{Y})$ , and a rule to choose one hypothesis over the other. The test statistic and the rule can be constructed with many different objectives in mind. The most common objectives are [28]:

- *Probability of false alarm* ( $P_{\text{FA}}$ ), which is the probability to incorrectly detect a change. The quantity  $1 - P_{\text{FA}}$  is denoted the *level* of the test.
- *Probability of detection* ( $P_{\text{D}}$ ), which is the probability of correctly detecting a change when a change has occurred. This property is also known as the *power* of a test.
- *Time to detect* ( $t_{\text{D}}$ ), which is the expected value of the time between the change and when it is detected.
- *Mean time between false alarms* ( $t_{\text{FA}}$ ), which is the expected value of the time between two false alarms.

The probability of false alarm and the probability of detection are important properties often used together to characterize a detector. Furthermore, the two are coupled; a given probability of false alarm,  $P_{\text{FA}}$ , limits how large the probability of detection,  $P_{\text{D}}$ , can be, and a given  $P_{\text{D}}$  puts a lower bound on  $P_{\text{FA}}$ . Due to the coupling,  $P_{\text{D}}$  is often plotted against  $P_{\text{FA}}$  in what is called a *receiver operating statistics* (ROC) diagram. The ROC diagram can

then be used to decide a reasonable compromise between the probability of detection and the probability of false alarm.

With the notation adopted in this thesis, a hypothesis test involves estimating, explicitly or implicitly, the parameter  $f_t$  in (6.1). The estimate is then used, often indirectly, in the test statistic to decide which hypothesis to choose. One approach, not necessarily based on stochastic theory, is to compute the least squares estimate of  $f_t$  and decide if it significantly differs from 0, or not. With better estimates of the parameter, e.g., the *best linear unbiased estimate* (BLUE), the *minimum variance estimate* (MVE), or the *maximum likelihood estimate* (MLE), it is often possible to derive tests that are more efficient. This opens up for nonlinear estimators such as the particle filter, which was shown to improve the quality of the estimate in Chapter 5. The particle filter also provides an estimate of the complete PDF of the estimated parameter, which can be utilized in the hypothesis test, see [46]. A general rule is that the more accurate the estimate of the parameter is, the better grounds for deciding between the hypotheses.

---

### Example 6.2: Hypothesis test

---

Let  $y \sim \mathcal{N}(f, 1)$ , where  $f$  equals 0 or  $\mu > 0$ , and design a hypothesis test for

$$\begin{cases} \mathcal{H}_0 : & \text{If } y \sim \mathcal{N}(0, 1), \text{ i.e., } f = 0, \\ \mathcal{H}_1 : & \text{If } y \sim \mathcal{N}(\mu, 1), \text{ i.e., } f = \mu. \end{cases}$$

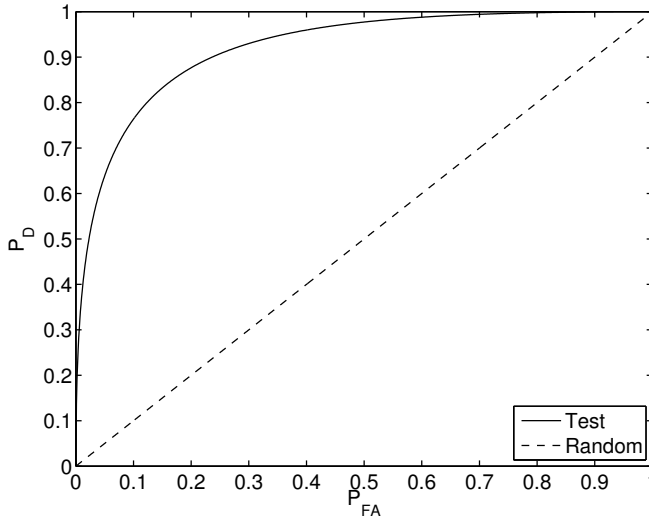
Let the test statistic be  $L(y) = y$  and decide for  $\mathcal{H}_0$  if  $y \leq \gamma$  and otherwise for  $\mathcal{H}_1$ . The probability of false alarm for the test is

$$P_{\text{FA}} = \Pr(y > \gamma | y \sim \mathcal{N}(0, 1)) = 1 - \Phi(\gamma)$$

where  $\Phi(\cdot)$  is the Gaussian CDF and the probability of detection is

$$P_{\text{D}} = \Pr(y > \gamma | y \sim \mathcal{N}(\mu, 1)) = 1 - \Phi(\gamma - \mu).$$

Hence, for  $\mu = 2$  and  $\gamma = 2$  the probability of false alarm is  $P_{\text{FA}} \approx 2\%$  and the probability of detection is  $P_{\text{D}} = 50\%$ . Varying  $\gamma$ , the ROC diagram in Figure 6.1 is obtained. Note the dashed line in the diagram that shows the performance obtained if a change is randomly detected with probability  $P_{\text{FA}}$ . Any test should be at least this good, or else it is better to randomly detect a change. Hence, it is always possible to obtain a test with  $P_{\text{D}} \geq P_{\text{FA}}$ .



**Figure 6.1:** ROC diagram for the detector in Example 6.2 with  $\mu = 2$ . The dashed line represents the result achieved for random detection.

## 6.2 Test Statistics

This section outlines the most commonly used test statistics: the *likelihood ratio* (LR) test, for deciding between simple hypotheses, and the *generalized likelihood ratio* (GLR) test, an extension to composite hypotheses of the regular likelihood ratio test.

### 6.2.1 Likelihood Ratio Test

For *simple* hypothesis tests, with only one alternative hypothesis, the *likelihood ratio test* [59] is one of the most well-known tests. The likelihood ratio test uses the ratio between the probabilities of the obtained measurements under the alternative hypothesis and the probability of the null hypothesis as test statistic,

$$L(\mathbb{Y}) = \frac{p(\mathbb{Y}|\mathcal{H}_1)}{p(\mathbb{Y}|\mathcal{H}_0)}. \quad (6.2)$$

The more likely the alternative hypothesis is compared to the null hypothesis, the larger  $L(\mathbb{Y})$  becomes, and *vice versa*. Therefore, if the test statistic is larger than a threshold,  $L(\mathbb{Y}) \geq \gamma$ , the null hypothesis,  $\mathcal{H}_0$ , is rejected, and if  $L(\mathbb{Y}) < \gamma$  the null hypothesis,  $\mathcal{H}_0$ , is accepted. The following notation will be used to represent this decision rule,

$$\begin{cases} \mathcal{H}_0, & \text{if } L(\mathbb{Y}) < \gamma \\ \mathcal{H}_1, & \text{if } L(\mathbb{Y}) \geq \gamma \end{cases} \iff L(\mathbb{Y}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma. \quad (6.3)$$

The probability of a false alarm with a given test statistic and the decision rule (6.3) is

$$P_{\text{FA}} = \Pr(L(\mathbb{Y}) \geq \gamma | \mathcal{H}_0) = \Pr\left(p(\mathbb{Y} | \mathcal{H}_1) \geq \gamma p(\mathbb{Y} | \mathcal{H}_0) \mid \mathcal{H}_0\right), \quad (6.4a)$$

and the probability of detection is

$$P_{\text{D}} = \Pr(L(\mathbb{Y}) \geq \gamma | \mathcal{H}_1) = \Pr\left(p(\mathbb{Y} | \mathcal{H}_1) \geq \gamma p(\mathbb{Y} | \mathcal{H}_0) \mid \mathcal{H}_1\right). \quad (6.4b)$$

The threshold  $\gamma$  determines how much more likely the alternative hypothesis must be before it is chosen over the null hypothesis. The larger the difference between the likelihoods, the easier it is to make a good decision. The difference can for instance be measured with the Kullback-Leibler information since the Kullback-Leibler information is a measure of how difficult it is to tell two distributions apart [54, 60].

---

**Example 6.3: Likelihood ratio test between Gaussians**

---

Assume a measurement  $y$  with

$$\begin{cases} \mathcal{H}_0 : & y \sim \mathcal{N}(0, 1) \\ \mathcal{H}_1 : & y \sim \mathcal{N}(\mu, 1). \end{cases}$$

The likelihood ratio test for deciding between  $\mathcal{H}_0$  and  $\mathcal{H}_1$  becomes

$$L(y) = \frac{p(y | \mathcal{H}_1)}{p(y | \mathcal{H}_0)} = \frac{\mathcal{N}(y; \mu, 1)}{\mathcal{N}(y; 0, 1)} = e^{\mu y - \mu^2/2} \underbrace{\frac{\mathcal{H}_1}{\mathcal{H}_0}}_{\geq \gamma}.$$

To find a suitable threshold,  $\gamma$ , to achieve a level,  $1 - P_{\text{FA}}$ , for the test, solve for  $\gamma$  in

$$P_{\text{FA}} = \Pr(L(y) \geq \gamma \mid \mathcal{H}_0) = \Pr\left(e^{\mu y - \frac{\mu^2}{2}} \geq \gamma \mid y \sim \mathcal{N}(0, 1)\right).$$

It is allowed to apply a logarithm to both sides of the inequality since  $\log(x)$  is strictly increasing for  $x > 0$ . Using the logarithm yields the simplified equation

$$P_{\text{FA}} = \Pr\left(\log(L(y)) \geq \log \gamma \mid \mathcal{H}_0\right) = \Pr\left(y \geq \underbrace{\frac{\log \gamma}{\mu} + \frac{\mu}{2}}_{=: \gamma'} \mid \mathcal{H}_0\right) = 1 - \Phi(\gamma'),$$

which is easily solved for  $y \sim \mathcal{N}(0, 1)$ . Once  $\gamma'$  is chosen, the probability of detection becomes

$$P_{\text{D}} = \Pr(y \geq \gamma' \mid \mathcal{H}_1) = 1 - \Phi(\gamma' - \mu).$$

Compare the test and the test statistic derived here with the identical result in Example 6.2. Hence, it is obvious that the test derived in Example 6.2 is indirectly based on the likelihood ratio test statistic.

---

To use the logarithm of the likelihood ratio, as in Example 6.3, often simplifies the involved expressions. This is for instance the case for distributions from the exponential family of distributions [72], of which the Gaussian distribution is a member.

## 6.2.2 Generalized Likelihood Ratio Test

The likelihood ratio test has one shortcoming, it requires both hypotheses to be simple to be applicable. In many situations this is overly restrictive, e.g., to determine if a parameter has its nominal value or not does in most situations result in a composite alternative hypothesis. When composite hypothesis are involved, the *generalized likelihood ratio* (GLR) test [48, 63, 64], defined as

$$L(\mathbb{Y}) = \frac{\sup_{f|\mathcal{H}_1} p(\mathbb{Y}|\mathcal{H}_1)}{\sup_{f|\mathcal{H}_0} p(\mathbb{Y}|\mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \quad (6.5)$$

can be used instead. Basically, the GLR test compares the highest likelihoods obtainable under the respective hypotheses. Implicitly, this is the same as estimating the parameter using a maximum likelihood estimator under the two hypotheses, and use the estimates to construct and use a regular likelihood ratio test.

In analogy with the likelihood ratio test, the probability of false alarm and the probability of detection are given by (6.4). However, the two suprema involved often considerably complicates the task of deriving explicit expressions.

## 6.3 Most Powerful Detector

What determines how useful a detector is depends on what it is supposed to be used for. Sometimes it is very important to have few false alarms,  $P_{FA} \ll 1$ , whereas in other situations it is very important not to miss any detections,  $P_D \gg 0$ . In yet other situations it may be important to detect faults fast, the time to detection  $t_d$  is short, and so forth. Unfortunately, most of these properties contradict each other, e.g., demanding few false alarms will inevitable make it more difficult to obtain a high detection rate. This chapter will from here on deal only with the relationship between  $P_{FA}$  and  $P_D$ , and optimality in terms of them. The following different definitions of optimality are commonly in use [5]:

- A *most powerful test* is a test that amongst all tests at a given level, equivalent to a given  $P_{FA}$ , maximizes the power,  $P_D$ , of the test.
- A *minimax test* is a test that minimizes the maximal risk of failing to detect any hypothesis.
- A *Bayes test* is a test that for a given *a priori* distribution for the hypotheses has the minimal probability of making an incorrect decision.

In this thesis the focus lies on optimality in terms of most powerfulness.

For testing between two simple hypotheses the Neyman-Pearson lemma, first introduced by Neyman and Pearson in the article series [65, 66], provides the statistics of the most powerful detector. The lemma is further elaborated on in [5, 59] and is therefore given without further explanation.

**Theorem 6.1 (Neyman-Pearson lemma)**

Every most powerful test between two simple hypotheses for a given probability of false alarm,  $P_{FA}$ , uses, at least implicitly, the test statistic

$$L(\mathbb{Y}) = \frac{p(\mathbb{Y}|\mathcal{H}_1)}{p(\mathbb{Y}|\mathcal{H}_0)},$$

and the decision rule

$$L(\mathbb{Y}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma,$$

where the threshold  $\gamma$  is obtained by solving for  $\gamma$  in

$$P_{FA} = \Pr(L(\mathbb{Y}) \geq \gamma \mid \mathcal{H}_0).$$

**Proof:** See proof of Theorem 1 in Chapter 3 of [59]. □

The Neyman-Pearson lemma (Theorem 6.1) does not only give a limit on how powerful tests that can be constructed. The theorem also states that the likelihood ratio test is the most powerful test for simple hypothesis, and that all most powerful tests will somehow be based on this test statistic.

For hypothesis tests including at least one composite hypothesis the situation is more complicated. One way to get an upper performance bound is to assume the true fault parameter to be known, and construct a simple hypothesis test based on this information. The test obtained under these conditions is called *clairvoyant*. If the simple clairvoyant test is optimal, its performance constitutes an upper bound for the performance of any detector working on the original problem.

The concept of most powerful tests is not general enough to handle composite tests because there may be different most powerful tests depending on the actual parameter value. The *uniformly most powerful* (UMP) property is therefore introduced to solve this problem. A test is UMP if it is most powerful for every possible parameter value under the alternative hypothesis.

**Definition 6.2.** A hypothesis test that is most powerful for all  $f$  under  $\mathcal{H}_1$  is said to be a *uniformly most powerful* (UMP) test.

The concept of UMP tests is the natural extension of most powerfulness to handle composite tests. The price paid for introducing the extra degree of freedom in the composite hypothesis is that it is much harder to find UMP tests than to find most powerful tests. Hence, there is no generalized Neyman-Pearson lemma for composite tests. In [59] examples are given both of cases when UMP tests exist and of cases when it is possible to show that no UMP test exists. For instance, there is no UMP test between  $\mathcal{H}_0 : f = f^0$  and  $\mathcal{H}_1 : f \neq f^0$  unless  $f$  is further restricted, or infinite information is available [59]. The latter case will be studied next.

## 6.4 Asymptotic Generalized Likelihood Ratio Test

Theorem 6.1 shows that the likelihood ratio test is optimal for simple hypotheses. Unfortunately, the strong optimality properties of the likelihood ratio test do not carry over

to composite hypothesis tests and the GLR test. In fact, the exact optimality properties of the GLR test are unknown. However, the GLR test is known to be optimal in several special cases [5]. One case when optimality can be shown is when asymptotic information is available. Therefore, the sequel of this chapter will be used to study the asymptotic properties of the GLR test.

### Theorem 6.2

*The generalized likelihood ratio test is asymptotically uniformly most powerful amongst all tests that are invariant (invariance imposes no restriction for the usual cases, see [59] for details), when the value of  $\mathbb{F}$ , the stacked faults, is the only thing that differs between the null hypothesis and the alternative hypothesis. Furthermore, the asymptotic test statistic is given by*

$$L'(\mathbb{Y}) := 2 \log L(\mathbb{Y}) \stackrel{a}{\sim} \begin{cases} \chi_{n_{\mathbb{F}}}^2, & \text{under } \mathcal{H}_0 \\ \chi_{n_{\mathbb{F}}}^{\prime 2}(\lambda), & \text{under } \mathcal{H}_1 \end{cases},$$

where  $L(\mathbb{Y})$  is the generalized likelihood ratio test statistic,  $\chi_n^{\prime 2}(\lambda)$  is the non-central  $\chi^2$  distribution of order  $n$  with non-centrality parameter  $\lambda$ , and

$$\lambda = (\mathbb{F}^1 - \mathbb{F}^0)^T \mathcal{I}(\mathbb{F} = \mathbb{F}^0)(\mathbb{F}^1 - \mathbb{F}^0).$$

Here  $\mathcal{I}(\mathbb{F} = \mathbb{F}^0)$  is the Fisher information of the stacked faults, at the true value  $\mathbb{F}^0$ , under  $\mathcal{H}_0$ , and  $\mathbb{F}^1$  is the true value of  $\mathbb{F}$  under  $\mathcal{H}_1$ .

**Proof:** See [48, Ch. 6]. □

For anything but theoretical argumentation, the assumption of infinite information is unrealistic. However, the asymptotic behavior constitutes a fundamental upper performance limitation and as such it can be compared to the Cramér-Rao lower bound for estimation (see Chapter 5). The asymptotic GLR properties can therefore be used to indicate how much better performance could be hoped for by utilizing available information about non-Gaussian noise even though no best linear detector comparable to the Kalman filter exists. Furthermore, in practice the GLR test usually performs quite well for moderately sized series of measurements [48]. Hence, the asymptotic behavior indicates what kind of performance to expect.

## 6.4.1 Wald Test

One way to motivate the asymptotic GLR test statistics is to optimally estimate the fault parameter,  $\mathbb{F}$ , and based on this estimate, with known statistics, do hypothesis testing. This approach is called the *Wald test* if the parameter is estimated with an MLE, and it is known that it has the same favorable asymptotic properties as the GLR test [48, 90].

To use the Wald test, the first step is to obtain an estimate,  $\hat{\mathbb{F}}$ , of the fault parameter. From Chapter 5, the optimal estimate is known to be asymptotically distributed according to

$$\hat{\mathbb{F}} \stackrel{a}{\sim} \mathcal{N}(\mathbb{F}, \mathcal{I}_{\mathbb{F}}^{-1}),$$

assuming, without loss of generality, that  $\mathbb{F}^0 = 0$ . Normalizing the estimate to obtain an estimate with unit covariance matrix yields

$$\hat{\mathbb{F}} := \mathcal{I}_{\mathbb{F}}^{\frac{1}{2}} \hat{\mathbb{F}} \stackrel{a}{\sim} \mathcal{N}(\mathcal{I}_{\mathbb{F}}^{\frac{1}{2}} \mathbb{F}, I).$$

Under  $\mathcal{H}_0$ , where  $\mathbb{F} = 0$ , the estimate is distributed according to

$$\hat{\mathbb{F}} \stackrel{a}{\sim} \mathcal{N}(\mathcal{I}_{\mathbb{F}}^{\frac{1}{2}} \mathbb{F}^0, I) = \mathcal{N}(0, I).$$

With  $L(\mathbb{Y}) = \|\hat{\mathbb{F}}\|_2^2$  as test statistic, the distribution of  $L(\mathbb{Y})$  under  $\mathcal{H}_0$  becomes

$$L(\mathbb{Y}) := \|\hat{\mathbb{F}}\|_2^2 \stackrel{a}{\sim} \chi_{n_{\mathbb{F}}}^2.$$

Under the alternative hypothesis,  $\mathcal{H}_1$ ,

$$\hat{\mathbb{F}} \stackrel{a}{\sim} \mathcal{N}(\mathcal{I}_{\mathbb{F}}^{\frac{1}{2}} \mathbb{F}^1, I)$$

yielding the test statistics distribution

$$L(\mathbb{Y}) = \hat{\mathbb{F}}^{1T} \mathcal{I}_{\mathbb{F}} \hat{\mathbb{F}}^1 \stackrel{a}{\sim} \chi_{n_{\mathbb{F}}}^{\prime 2}(\lambda),$$

where  $\mathbb{F}^{1T}$  is a shorthand notation for  $(\mathbb{F}^1)^T$  and  $\lambda = \mathbb{F}^{1T} \mathcal{I}_{\mathbb{F}} \mathbb{F}^1$ . With this information a suitable threshold can be found for the decision rule

$$L(\mathbb{Y}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma'.$$

The test statistics derived this way should be recognized as the same as for the asymptotic GLR test in Theorem 6.2. However, the test statistics are only valid under the asymptotic assumption, and the behavior of the Wald test and the GLR test may differ for finite information.

The derivation of the Wald test indicates how an approximation of the system noise affects the performance of a test. With a Gaussian approximation of the noise, the variance of the estimate of the fault parameter increases (see Chapter 5), which makes it more difficult to detect a fault.

## 6.4.2 Detection Performance

For the asymptotic GLR test, and consequently the asymptotic Wald test, with the threshold  $\gamma'$  the probability of a false alarm,  $P_{\text{FA}}$ , is

$$P_{\text{FA}} = \mathcal{Q}_{\chi_{n_{\mathbb{F}}}^2}(\gamma'), \quad (6.6)$$

where  $\mathcal{Q}_{\star}$  denotes the complementary cumulative distribution function of the distribution  $\star$ . This follows directly from the test statistics given by Theorem 6.2. Note that  $P_{\text{FA}}$  depends only on the choice of threshold,  $\gamma'$ , and not on  $\mathbb{F}^0$  or  $\mathbb{F}^1$ . The probability of detection is

$$P_{\text{D}} = \mathcal{Q}_{\chi_{n_{\mathbb{F}}}^{\prime 2}(\lambda)}(\gamma'), \quad (6.7)$$

where  $\lambda$  is defined in Theorem 6.2. The function  $\mathcal{Q}_{\chi^2_{n_{\mathbb{F}}}}(\lambda)(\gamma')$  is monotonously increasing in  $\lambda$  (moving the mean to the right lessens the risk that a detection is missed) thus any increase in  $\lambda$  will improve  $P_D$ . Note, it follows immediately that if the magnitude of  $\mathbb{F}^1$  increases it is easier to detect the change, and that if  $\mathcal{I}_{\mathbb{F}}$  increases  $P_D$  increases, *i.e.*, any non-Gaussian noise component will increase the probability of detection with preserved  $P_{FA}$ , unless the system suppresses the informative noise direction.

#### Example 6.4

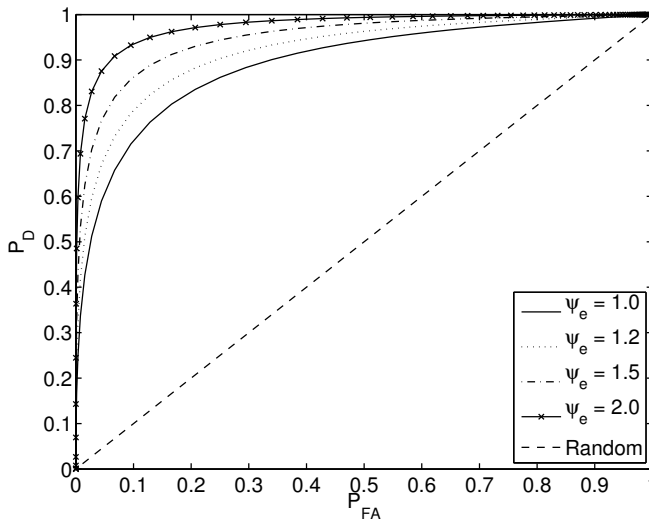
Consider measurements from

$$y_t = f + e_t, \quad t = 1, \dots, L, \quad (6.8)$$

where  $e_t$  is any noise with  $\text{var}(e_t) = 1$  and relative accuracy  $\Psi_e$ . It then follows that the normalized estimate of  $f$  is

$$\hat{f} \stackrel{a}{\sim} \mathcal{N}\left(\sqrt{\Psi_e L} f, 1\right),$$

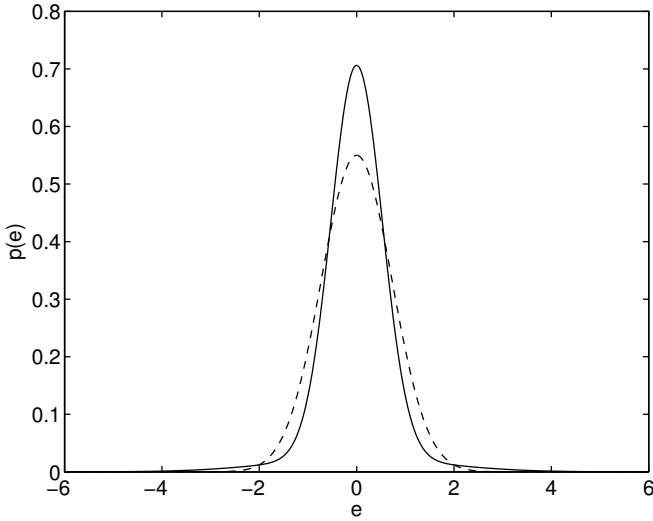
and subsequently to  $\lambda = \Psi_e L (f^1)^2 = \Psi_e L$  in the GLR statistics under the assumption  $f^1 = 1$ . The improved detection performance is illustrated with a ROC diagram in Figure 6.2. From Figure 6.2 it is clear that there is potentially much to be gained from utilizing information about non-Gaussian noise in this simple model, especially for small  $P_{FA}$  where the relative increase in  $P_D$  is substantial.



**Figure 6.2:** ROC diagram for (6.8) with  $n_{\mathbb{F}} = 1$ ,  $L = 5$  measurements, and  $\mathbb{F}^1 = 1$  for different values of  $\Psi_e$ . Random denotes what happens if all available information is discarded and a change is signaled randomly with probability  $P_{FA}$ .

To show the performance actually obtainable, assume that the measurement noise is subject to outliers,

$$e_t \sim 0.9\mathcal{N}(0, R) + 0.1\mathcal{N}(0, 10R), \quad (6.9)$$



**Figure 6.3:** Solid line, PDF for measurement noise with outliers, (6.9),  $\text{var}(e_t) = 1$ , and  $\Psi_e = 1.5$ . Dashed line shows the Gaussian approximation.

where  $R = 0.28$ , so that  $\text{var}(e_t) = 1$ . This is an instance of the outlier noise described in Section 2.3.2 with  $\mathcal{I}_e = \Psi_e = 1.5$ . The PDF is shown in Figure 6.3.

Simulations with this setup, using 10 000 Monte Carlo simulations and a numeric GLR test implementation, yields the ROC diagram in Figure 6.4. The result is promising since the simulations seem to come close to the performance bound (*cf.* Figure 6.2).

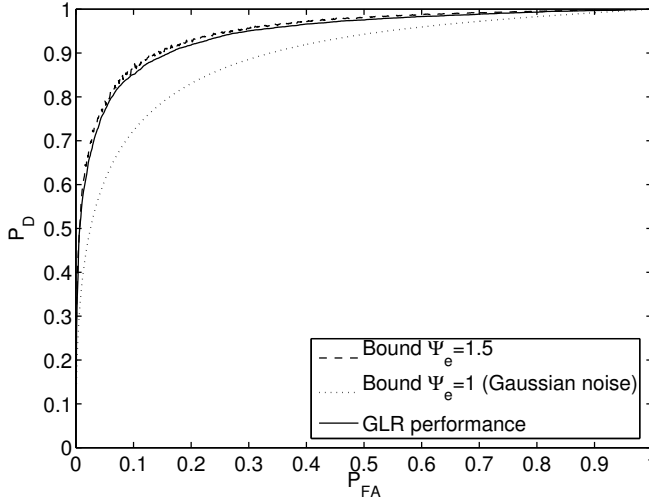
## 6.5 Uniformly Most Powerful Test for Linear Systems

In this section, the theory for general systems, presented in the previous sections, is applied to two problems with linear *residuals*. The section is divided into two main parts. First general results are derived for linear residuals and it is then shown how the theory can be applied to two special cases of change detection with linear systems. The reader only interested in the general results may skip the special cases and continue with the simulations in the next section.

### 6.5.1 Linear System Residuals

Most change detection methods do not deal directly with the measurements. The difference between the measurements and the measurements expected under the null hypothesis,

$$r_t = y_t - h(x_t, f_t^0), \quad (6.10)$$



**Figure 6.4:** ROC diagram for (6.8) with measurements (6.9) ( $\Psi_e = 1.5$ ) for 10 000 MC simulations. Optimal performance for  $\Psi_e = 1$  and  $\Psi_e = 1.5$  are found in Figure 6.2 as reference.

is studied instead. The difference  $r_t$  is called a *residual*. Without noise the residual should be zero under  $\mathcal{H}_0$ , but since noise is always present the residual should be small and average to zero over time if no fault is present. The exact residual distribution is determined by the system and the noise affecting it. For linear models with the fault structure discussed in Section 3.1.3 the residual is given by

$$r_t = y_t - H_t x_t = H_t^w w_t + e_t + H_t^f f_t = \underbrace{\sum_i H_t^f H^{f,i} \varphi_t^T \theta}_{=: H_t^\theta} + H_t^w w_t + e_t,$$

where  $\theta$  is the fault parameter and  $\varphi_t$  make up an basis to describe incipient faults. If  $L$  measurements in a time window are gathered, the stacked model becomes

$$\begin{aligned} \mathbb{R}_t &= \mathbb{Y}_t - \mathcal{O}_t x_{t-L+1} = \bar{H}_t^f \mathbb{F}_t + \bar{H}_t^w \mathbb{W}_t + \mathbb{E}_t \\ &= \bar{H}_t^\theta \theta + \underbrace{(\bar{H}_t^w \quad I)}_{=: \bar{H}_t^v} \underbrace{\begin{pmatrix} \mathbb{W}_t \\ \mathbb{E}_t \end{pmatrix}}_{=: \mathbb{V}_t} = \bar{H}_t^\theta \theta + \bar{H}_t^v \mathbb{V}_t, \end{aligned} \quad (6.11)$$

which is a linear regression in the fault parameter  $\theta$ , and where  $\mathbb{V}$  combines the noise effects over the window. Note that the dimension of  $\theta$  is not increased due to the stacking. This is significant because a consequence of this is that if the studied window is enlarged the information available about the elements in  $\theta$  increases.

Using the residual (6.11), the statistics of the asymptotic GLR test becomes

$$P_{\text{FA}} = \mathcal{Q}_{\chi_{n_\theta}^2}(\gamma) \quad (6.12a)$$

$$P_{\text{D}} = \mathcal{Q}_{\chi_{n_\theta}^{\prime 2}(\lambda)}(\gamma), \quad (6.12b)$$

where  $\gamma$  is the detection threshold and

$$\lambda = \theta^{1T} \bar{H}_t^{\theta T} \mathcal{I}_{\bar{H}_t^v \mathbb{V}} \bar{H}_t^\theta \theta^1 = \theta^{1T} \bar{H}_t^{\theta T} (\bar{H}_t^v \mathcal{I}_{\mathbb{V}}^{-1} \bar{H}_t^{vT})^{-1} \bar{H}_t^\theta \theta^1.$$

From this it is clear that if the intrinsic accuracy of  $\mathbb{V}_t$  is improved, *i.e.*, by introducing non-Gaussian components in the process noise or the measurement noise, the probability of detection increases. If the improvement is in a direction ignored by the system, the potential gain is lost.

It is sometimes useful to use only part of the residual derived above. This can be achieved by applying a projection when constructing  $\mathbb{R}_t$ , this way removing a subspace of the residual space known to be difficult to estimate or otherwise unsuitable for usage in the test. The parity space formulation, discussed below as a special case of this theory, uses a carefully chosen projection to remove the influence of the unknown initial state. Introducing a projection only changes the matrices  $\bar{H}_t^\theta$  and  $\bar{H}_t^v$  in the analysis.

## 6.5.2 Prior initial state knowledge

One of the unknown components of the residual description (6.11) is the initial state of the studied window,  $x_{t-L+1}$ . If an estimate of the state is available,  $\hat{x}_{t-L+1}$ , from measurements outside the window it can be used. However, an estimate is just an approximation of the true state and an extra noise term is introduced in the residual description. The residual becomes,

$$\begin{aligned} \mathbb{R}_t &= \mathbb{Y}_t - \mathcal{O}_t \hat{x}_{t-L+1} = \mathcal{O}_t \tilde{x}_{t-L+1} + \bar{H}_t^w \mathbb{W}_t + \mathbb{E}_t + \bar{H}_t^f \mathbb{F}_t \\ &= \bar{H}_t^\theta \theta + (\mathcal{O}_t \quad \bar{H}_t^w \quad I) \begin{pmatrix} \tilde{x}_{t-L+1} \\ \mathbb{W}_t \\ \mathbb{E}_t \end{pmatrix}, \end{aligned} \quad (6.13)$$

where  $\tilde{x}_{t-L+1} := x_{t-L+1} - \hat{x}_{t-L+1}$ . The distribution of  $\tilde{x}_{t-L+1}$  depends on the estimator used to obtain  $\hat{x}_{t-L+1}$ . Any estimate of  $x_{t-L+1}$  will do, but with a better estimate the detection performance improves. Assuming that  $\hat{x}_{t-L+1}$  is not based on studied data in the window avoids dependencies between the noise terms, which may be difficult to handle. The residual formed this way still fits into the general linear regression framework given by (6.11), hence the analysis in Section 6.5.1 still applies and the test statistics are:

$$P_{\text{FA}} = \mathcal{Q}_{\chi_{n_\theta}^2}(\gamma) \quad (6.14a)$$

$$P_{\text{D}} = \mathcal{Q}_{\chi_{n_\theta}^{\prime 2}(\lambda)}(\gamma), \quad (6.14b)$$

where

$$\lambda = \theta^{1T} \bar{H}_t^{\theta T} \left( (\mathcal{O} \quad \bar{H}_t^w \quad I)^T \mathcal{I}_{\mathbb{V}_t}^{-1} (\mathcal{O} \quad \bar{H}_t^w \quad I) \right)^{-1} \bar{H}_t^\theta \theta^1,$$

with  $\mathbb{V}_t = \begin{pmatrix} \tilde{x}_{t-L+1} \\ \mathbb{W}_t \\ \mathbb{E}_t \end{pmatrix}$ .

### 6.5.3 Parity Space

Sometimes it is favorable to look only at contributions to the residual that could not come from the nominal system. To study only this part of the residual is called to use *parity space* or *analytic redundancy* [5, 25]. One reason to work in parity space is that no estimate of the initial state is available, or that the available estimate is poor and unreliable. The residual are obtained in parity space as

$$\begin{aligned}\mathbb{R}_t &= \mathcal{P}_\mathcal{O}^\perp (\mathbb{Y}_t - \mathcal{O}_t x_{t-L+1} - \bar{H}_t^u \mathbb{U}_t) = \mathcal{P}_\mathcal{O}^\perp \mathbb{Y}_t - \mathcal{P}_\mathcal{O}^\perp \bar{H}_t^u \mathbb{U}_t \\ &= \mathcal{P}_\mathcal{O}^\perp (\bar{H}_t^w \mathbb{W}_t + \mathbb{E}_t + \bar{H}_t^f \mathbb{F}_t) = \mathcal{P}_\mathcal{O}^\perp \bar{H}_t^\theta \theta + \mathcal{P}_\mathcal{O}^\perp (\bar{H}_t^w \quad I) \begin{pmatrix} \mathbb{W}_t \\ \mathbb{E}_t \end{pmatrix},\end{aligned}\quad (6.15)$$

where  $\mathcal{P}_\mathcal{O}^\perp$  is a projection matrix such that  $\mathcal{P}_\mathcal{O}^\perp \mathcal{O} = 0$ , i.e., a projection into the complement of the state space. This way any contribution from  $x_{t-L+1}$  is effectively removed from the residuals, and hence it is unimportant that  $x_{t-L+1}$  is unknown.

The projection  $\mathcal{P}_\mathcal{O}^\perp$  can be constructed using a *singular value decomposition* (SVD) [26] of the extended observability matrix  $\mathcal{O}$ . First, take the SVD of  $\mathcal{O}$ ,

$$\mathcal{O} = U \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} V^T = (U_1 \quad U_2) \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} (V_1 \quad V_2)^T, \quad (6.16)$$

where  $U$  and  $V$  are unitary matrices, split into blocks so that  $U_1$  spans the range of  $\mathcal{O}$  and  $V_2$  the null space of  $\mathcal{O}$ , and  $\Sigma$  is a diagonal matrix with the non-zero singular values of  $\mathcal{O}$ . After that, choose  $\mathcal{P}_\mathcal{O}^\perp = U_2^T$  to get a projection with the desired properties. Note,

$$\text{cov}(\mathbb{R}_t) = \mathcal{P}_\mathcal{O}^\perp \bar{H}_t^w \Sigma_{\mathbb{W}_t} \bar{H}_t^{wT} \mathcal{P}_\mathcal{O}^{\perp T} + \mathcal{P}_\mathcal{O}^\perp \Sigma_{\mathbb{E}_t} \mathcal{P}_\mathcal{O}^{\perp T}, \quad (6.17)$$

and  $\mathcal{P}_\mathcal{O}^\perp$  has full row rank and therefore  $\text{cov}(\mathbb{R}_t) \succ 0$  if  $\text{cov}(\mathbb{E}_t) \succ 0$ . Furthermore, in order for the parity space to contain any interesting information the system must be such that  $\text{rank}(\mathcal{P}_\mathcal{O}^\perp \bar{H}_t^f) > 0$  in order for all faults to show up in the residuals.

In the parity space setting the signal part of the model is removed using a projection leaving only  $\mathcal{I}_\mathbb{W}$  and  $\mathcal{I}_\mathbb{E}$  as interesting quantities. The statistics are

$$P_{\text{FA}} = \mathcal{Q}_{\chi_{n_\theta}^2}(\gamma) \quad (6.18a)$$

$$P_{\text{D}} = \mathcal{Q}_{\chi_{n_\theta}^{\prime 2}(\lambda)}(\gamma), \quad (6.18b)$$

where

$$\lambda = \theta^{1T} \bar{H}_t^{\theta T} \mathcal{P}_\mathcal{O}^{\perp T} \left( (\mathcal{O} \quad \bar{H}_t^w \quad I)^T \mathcal{I}_{\mathbb{V}_t}^{-1} (\bar{H}_t^w \quad I) \right)^{-1} \mathcal{P}_\mathcal{O}^\perp \bar{H}_t^\theta \theta^1$$

with  $\mathbb{V}_t = \begin{pmatrix} \mathbb{W}_t \\ \mathbb{E}_t \end{pmatrix}$ .

## 6.6 Applications of the Theory

The treatment of change detection in this chapter is concluded with several simulation studies. The same linear systems as in Chapter 5 are used; the constant velocity model and the DC motor. Both systems will be affected by non-Gaussian process noise and

measurement noise. Using the same systems as in the estimation simulations allows for parallels to be drawn between detection and estimation performance. Furthermore, since the systems are thoroughly described in Section 5.4 it is not necessary to do that again, therefore the descriptions of the systems in this section focuses on the faults to detect.

Detection is in all simulations performed in parity space, on a time window of  $L = 6$  measurements. Parity space is used to, as far as possible, separate the detection problem from the estimation problem studied in Chapters 4 and 5. Working in parity space eliminates the  $x_{t-L+1}$  term that would otherwise have to be estimated and that way impose properties of the estimate on the detection performance. Stacked versions of linear models are described in Section 3.1.4 and the projection down into the parity space using the method given in Section 6.5.3. The result is a residual description on the form (6.15).

GLR tests are used to detect faults in stacked residual formulations of the systems. All statistical properties of the stacked model are known, and it is therefore possible to derive the analytical likelihood ratio if the fault is given. However, the fault is assumed unknown and the analytic MLE of the fault is difficult to obtain if the noise is non-Gaussian. Therefore, the MLE is computed numerically resulting in an approximate GLR test, which seems to perform quite well in practice.

For comparison, a GLR test has also been designed in each case assuming that the noise is Gaussian, *i.e.*, all noise distributions have been assumed Gaussian for the design of the GLR test. The detector derived in this way is used for comparison. However, note that there are no results similar to the BLUE to support this approach, nonetheless this is a result that could be achieved if non-Gaussian effects are ignored.

### 6.6.1 Constant Velocity Model

The first simulation uses the constant velocity model (see Section 5.4.1). This time the state of the system is of minor interest, and the main objective is to determine if the system is affected by a fault or not. The potential fault enters the system the same way the process noise does. The model therefore becomes

$$x_{t+1} = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix} x_t + \begin{pmatrix} \frac{1}{2}T^2 \\ T \end{pmatrix} w_t + \begin{pmatrix} \frac{1}{2}T^2 \\ T \end{pmatrix} f_t \quad (6.19a)$$

$$y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} x_t + e_t, \quad (6.19b)$$

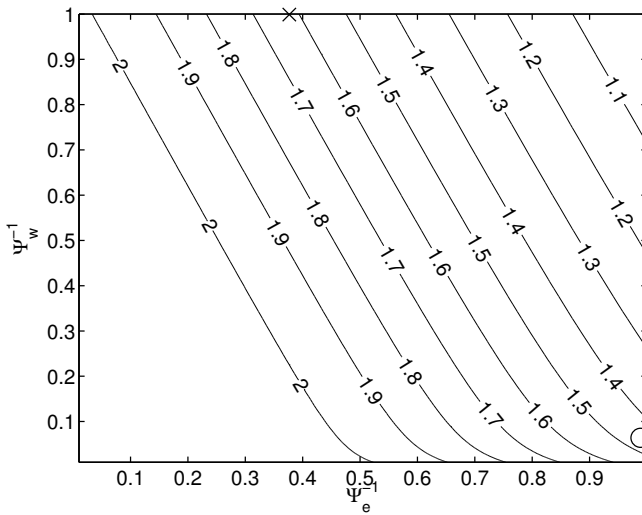
with  $T = 1$  and  $f_t$  as the fault term. The process noise and measurement noise are known to be mutually independent white noise with  $\text{var}(w_t) = \text{var}(e_t) = 1$ . To lower the complexity of the detection problem the fault is assumed to be either present or not, *i.e.*,  $f_t = \theta$  without any incipient behavior. Throughout the simulations the true fault profile is

$$f_t = \begin{cases} 0, & t \leq 25 \\ 1.5, & t > 25, \end{cases}$$

*i.e.*, the fault appears suddenly after time  $t = 25$ . In terms of the constant velocity model this can be interpreted as an unexpected maneuver, or that the measurement platform has started to move.

The optimal detection performance in terms of the asymptotic GLR performance for the constant velocity model can be computed from the test statistics in (6.18), given  $L =$

6,  $f_t^1 = 1.5$ , and the noise characteristics. Figure 6.5 shows the probability of detection, if  $P_{FA} = 1\%$ , as a function of the intrinsic accuracy of system noise. The probability of detection has been normalized to be 1 for Gaussian noise (for which  $P_D = 48\%$ ), to simplify the comparison of the detection performance. That is, if the contour plot indicates detection performance 1.5 for a given noise characteristics,  $P_D$  is 50% higher with that noise than with Gaussian noise. In this way, Figure 6.5 is similar to Figure 5.1, both display how much can potentially be gained using all noise information available. Note that with a different  $P_{FA}$  or magnitude of the fault, a different result is obtained due to the nonlinearities in the expressions, but the main characteristics of the result will be the same. Figure 6.5 shows that for non-Gaussian noise it is asymptotically possible to



**Figure 6.5:** Relative probability of detection for the given constant velocity model at  $P_{FA} = 1\%$  for detection in parity space and a window size  $L = 6$ . In the contour plot 1 corresponds to  $P_D = 48\%$  which is obtained for Gaussian noise.

gain detection performance. Furthermore, the performance increases faster in the relative accuracy of the measurement noise than in the relative accuracy of the process noise. This indicates that describing the measurement noise correctly is more important for the performance than the describing process noise.

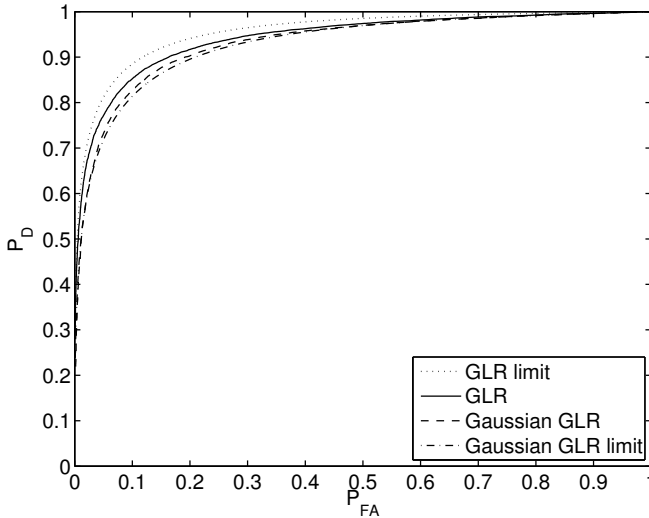
### Bimodal Measurement Noise

The first simulation, using the constant velocity model, features Gaussian process noise,  $\Psi_w = 1$ , and the non-Gaussian measurement noise

$$e_t \sim \frac{9}{10}\mathcal{N}(0.2, 0.3) + \frac{1}{10}\mathcal{N}(-1.8, 3.7),$$

with  $\Psi_e = 2.7$ . The detection performance achieved for 1 000 Monte Carlo simulations is given as a ROC diagram in Figure 6.6. The ROC diagram shows that by utilizing the extra information available in the non-Gaussian measurement noise the probability of detection

is improved, however not as much as indicated by the asymptotic results. The same system was used in Section 6.6.1 for a Monte Carlo study of estimation performance, with a similar result. Furthermore, the performance obtained when assuming Gaussian noise is better than expected for large  $P_{FA}$ . For  $P_{FA}$  close to 1, it is difficult to tell the difference between the approximative test and the test based on correct statistics. Hence, the behavior when using a Gaussian approximation is not as easy to predict as for estimation, where the BLUE can be used.



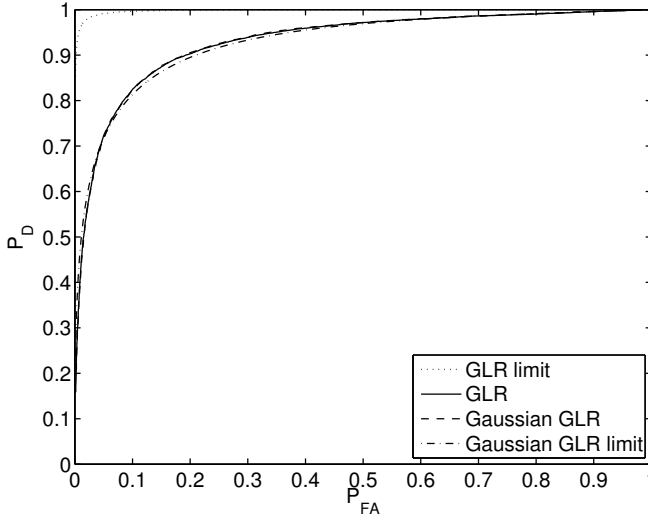
**Figure 6.6:** ROC diagram derived from 1 000 Monte Carlo simulations for the parity space formulation of the constant velocity model with bi-Gaussian measurement noise,  $L = 6$ , and the fault  $f^1 = 1.5$ .

### Tri-Gaussian Process Noise

For the second simulation, with the constant velocity model, where  $e_t$  is kept Gaussian,  $\Psi_e = 1$ , and

$$w_t \sim 0.075\mathcal{N}(-2.5, 0.065) + 0.85\mathcal{N}(0, 0.065) + 0.075\mathcal{N}(+2.5, 0.065),$$

with  $\Psi_w = 15.4$ . This is the same system as was used in Section 5.4.1, where the discouraging result was that no improved performance could be found using a particle filter. The detection results are similar, see Figure 6.7. No difference can be observed between the detection result based on a correct noise description and the test using the Gaussian approximation. It is however in line with what could be expected since if the estimate cannot be improved then there is little new information to base a better detector on.



**Figure 6.7:** ROC diagram derived from 1 000 Monte Carlo simulations of a parity space formulation of the constant velocity model with tri-Gaussian process noise,  $L = 6$ , and the fault  $f^1 = 1.5$ .

## 6.6.2 DC Motor

The second system used for simulations is the DC motor, see Section 5.4.2. The DC motor is affected by a load disturbance,  $f_t$ , according to the model

$$x_{t+1} = \begin{pmatrix} 1 & 1 - e^{-1} \\ e^{-1} & 0 \end{pmatrix} x_t + \begin{pmatrix} e^{-1} \\ 1 - e^{-1} \end{pmatrix} w_t + \begin{pmatrix} e^{-1} \\ 1 - e^{-1} \end{pmatrix} f_t \quad (6.20a)$$

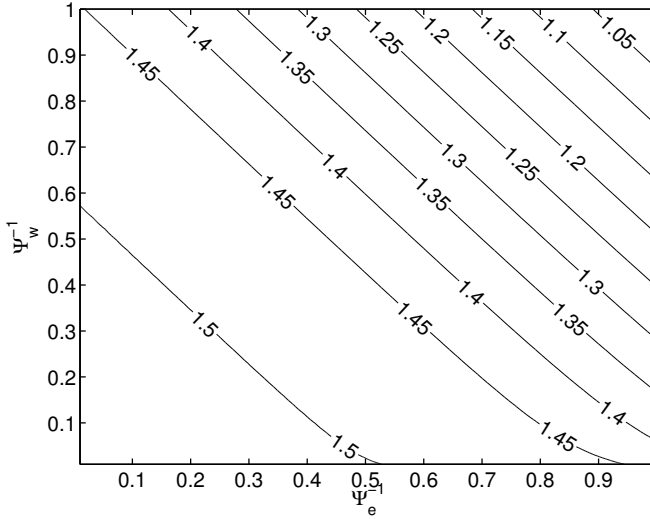
$$y_t = (1 \ 0) x_t + e_t, \quad (6.20b)$$

where  $w_t$  and  $e_t$  are independent but not necessarily Gaussian noise. This fault can be caused by a load being suddenly added to the drive shaft, a bad bearing in the motor, or an offset in the voltage being used to control the motor. It is here assumed that the load is either there or not, *i.e.*,  $f_t = \theta$ . In the Monte Carlo simulations the fault has the profile

$$f_t = \begin{cases} 0, & t \leq 25 \\ \frac{\pi}{90}, & t > 25 \end{cases}.$$

That is, 25 time units into the experiment something happens to the DC motor. The motor starts to accelerate  $\frac{\pi}{90}$  more than expected, *e.g.*, due to an offset in the control input.

As for the constant velocity model, a contour plot of the relative probability gain for  $P_{FA} = 1\%$  is given in Figure 6.8 (*cf.* the relative estimation performance in Figure 5.4). Compared to Figure 6.5 the relative gain obtainable according to Figure 6.8 is smaller, however this is in part because the detectability for Gaussian noise here is greater and the possibility to increase it is limited.



**Figure 6.8:** Relative probability of detection for the given constant velocity model at  $P_{FA} = 1\%$  for detection in parity space and a window size  $L = 6$ . In the contour plot 1 corresponds to  $P_D = 67\%$  which is obtained for Gaussian noise.

**Measuremens with Outliers**

The first simulation study with the DC motor involves Gaussian process noise and outliers in the measurements (cf. Section 5.4.2),

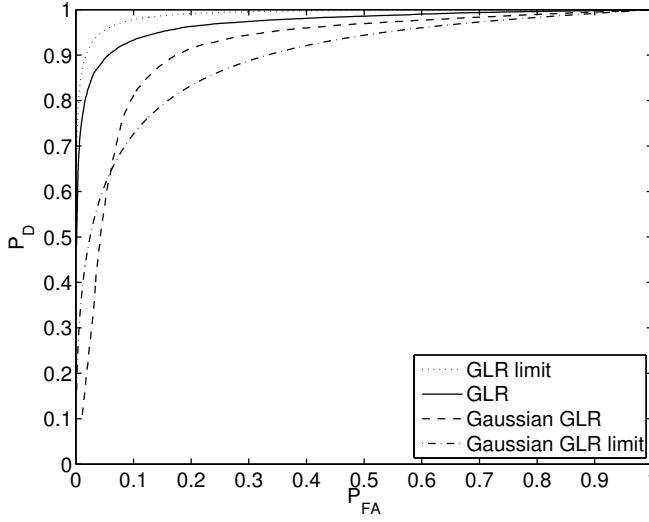
$$w_t \sim \mathcal{N}\left(0, \left(\frac{\pi}{180}\right)^2\right)$$

$$e_t \sim \mathcal{N}_2\left(\left(0.9, 0, \left(\frac{\pi}{180}\right)^2\right), \left(0.1, 0, \left(\frac{10\pi}{180}\right)^2\right)\right),$$

with  $\Psi_w = 1$  and  $\Psi_e = 9$ . The contour plot in Figure 6.8 indicates, even though it is not exactly the correct noise levels, that detection performance could improve substantially if the outliers are correctly handled in the test.

The result of the simulations is presented in the ROC diagram in Figure 6.9. Especially for low false alarm rates,  $P_{FA} \ll 1$ , the gain is substantial and close to what is possible to achieve with an asymptotic GLR test. Considering detection in terms of the Wald test the good detection performance is not surprising since it was shown in Section 5.4.2 that the estimation performance for this setup is close to the CRLB.

Once again, the GLR statistics under the false assumption that all noise is Gaussian prove to be a rather poor description of the behavior of what happens when non-Gaussian components are neglected. Furthermore, Figure 6.9 clearly shows for low false alarm rates, the asymptotic Gaussian GLR test is better than the performance achieved, whereas for larger  $P_{FA}$ , the approximative test outperforms the asymptotic Gaussian GLR test.



**Figure 6.9:** ROC diagram derived from 1 000 Monte Carlo simulations of a parity space formulation of the DC motor with outliers in the measurements,  $L = 6$ , and the fault  $f^1 = \frac{\pi}{90}$ .

### Load Disturbances

Finally, the DC motor is studied with Gaussian measurement noise,  $\Psi_e = 1$ , and bimodal process noise,  $\Psi = 17$ ,

$$w_t \sim \mathcal{N}_2 \left( \left( 0.8, 0, \left( \frac{\pi}{180} \right)^2 \right), \left( 0.2, \frac{-10\pi}{180}, \left( \frac{\pi}{180} \right)^2 \right) \right)$$

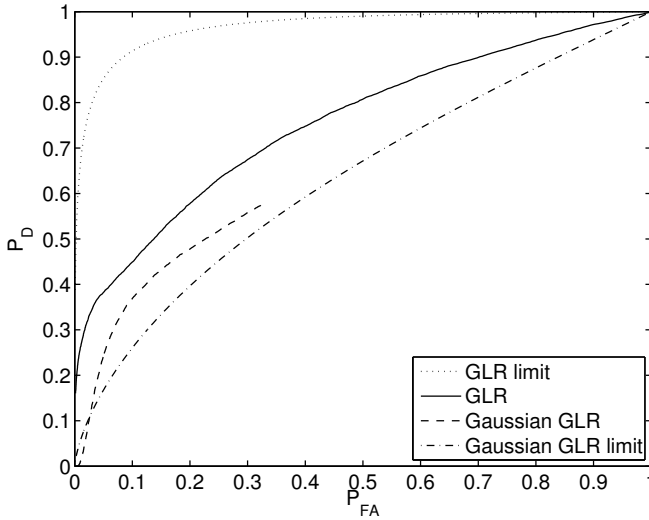
$$e_t \sim \mathcal{N} \left( 0, \left( \frac{\pi}{180} \right)^2 \right).$$

The same system was studied in Section 5.4.2, that time with respect to estimation performance. It was in Section 5.4.2 shown that with knowledge about the true statistics of the noise it was possible to improve estimation performance. Based on the previous detection experiments, and the indication in Figure 6.8 that the detection performance could be improved using the actual noise distributions for the test statistic, an improvement is to be expected.

The result of 1 000 Monte Carlo simulations is found as a ROC diagram in Figure 6.10. The performance using the correct noise distributions is clearly better than the performance achieved approximating all distributions with Gaussian noise. Note that the complete ROC curve is not obtained for the test based on the Gaussian approximation, due to numerical errors when computing  $L(\mathbb{Y}_t)$  for the approximative test. Once again, this shows the difficulty involved in treating non-Gaussian noise as Gaussian noise.

### 6.6.3 Observations

The results from the simulations conducted both with the constant velocity model and the DC motor can be summarized with:



**Figure 6.10:** ROC diagram derived from 1000 Carlo simulations of a parity space formulation of the DC motor with bi-Gaussian process noise,  $L = 6$ , and the fault  $f^1 = \frac{\pi}{90}$ .

- Studying the asymptotic GLR performance, it is potentially much to be gained from taking care of non-Gaussian effects.
- It is not enough to study the asymptotic GLR test performance for a system with Gaussian noise to predict the effects of a Gaussian approximation. The simulations indicate that this approach is overly optimistic for small  $P_{FA}$ , whereas it for large  $P_{FA}$  is more pessimistic than it needs to be. With this in mind, the asymptotic GLR test with Gaussian noise can be used to hint the behavior of an approximative detector.
- It is in the conducted experiments more difficult to extract information from non-Gaussian process noise compared to non-Gaussian measurement noise. Similar results were obtained for the estimation problem in Section 5.4.3.
- The simulations indicate that it is for small  $P_{FA}$  the most is gained by using a correct noise description.



# 7

---

## Concluding Remarks

IN THE INTRODUCTION to this thesis it is described how linear and Gaussian system approximations are often used for estimation and detection applications. The implications of the approximations are seldom analyzed, even though methods to handle the true systems exist. This chapter concludes the work in this thesis aimed towards giving guidelines for when noise approximations are acceptable in linear systems. Based on the results obtained ideas of future work are given.

### 7.1 Conclusions

The work in this thesis deals with linear systems with non-Gaussian noise from an estimation and detection perspective. For this class of models it is shown how the information content in the system noise affects the performance obtained with approximative methods. For this purpose *intrinsic accuracy* and *relative accuracy* are used. A method is given to compute the intrinsic accuracy, and it is shown that Gaussian mixtures, that to the naked eye look Gaussian, may be much more informative than a Gaussian approximation. If the intrinsic accuracy is large this indicates that a Gaussian noise approximation is unfavorable.

To determine if a noise approximation affects estimation performance the *Cramér-Rao lower bound* (CRLB) can be studied. Comparing the CRLB to the performance with linear systems obtained with the *Kalman filter*, which is known to be the *best linear unbiased estimator* (BLUE), it is possible to determine how much can be gained with a correct noise description in combination with a nonlinear filter, *e.g.*, a particle filter. The gain depends on properties of the system, if the relative accuracy of the system noise is high disregarding non-Gaussian properties tend to affect the estimation performance. Monte Carlo simulations verify this, but also show that the CRLB analysis sometimes must be combined with other methods to get conclusive results since the lower bound is not always reached.

Detection performance is discussed in terms of the uniformly most powerful asymptotic *generalized likelihood ratio* (GLR) test, for which the intrinsic accuracy plays an important role. If the performance utilizing the true noise distributions is much better than what is achieved for Gaussian noise, approximations should be avoided. Within the framework derived for detection in linear systems it is shown, using Monte Carlo simulations, that the detection improvement may be substantial, and also that it is difficult to predict the properties of a detector based on approximative noise distributions.

Based on results derived in this thesis it is possible to indicate if a Gaussian approximation is acceptable or not for a linear estimation or detection problem, given a complete system description including noise properties. However, work remains to be done to strengthen the theory to give more conclusive answers.

## 7.2 Further Work

Some questions still remain to be answered to get a complete framework to handle estimation and detection in nonlinear non-Gaussian systems in an satisfactory manner. Some issues regard linear systems:

- What performance is actually obtainable? The CRLB is a lower bound, but it is not always reached as seen in simulations. The situation is similar for the asymptotic GLR statistics. Is it possible to tell if/how optimal performance is reached?
- What about robustness? How sensitive are the results to errors in the model, both in the system dynamics and in the noise distributions?
- Are there better measures than the CRLB and the asymptotic GLR statistics to describe optimal performance? If so, which properties of the system are then interesting?

Furthermore, what approach is best for extending the framework to nonlinear systems? In principle, the techniques in the thesis can be used to obtain optimality results for nonlinear systems, however it is difficult predict the properties of different approximative methods, as seen in the study of the bearings-only measurements. Another problem is that results derived for general nonlinear systems tend to be less intuitive than the results for linear system, and results are often difficult and expensive to compute. All these are questions it would be interesting to continue working with.

# Appendix





---

# Notational Conventions

## Abbreviations and Acronyms

Abbreviation	Meaning
AFMM	Adaptive forgetting through multiple models
BLUE	Best linear unbiased estimate/estimator
CDF	Cumulative distribution function
CUSUM	Cumulative sum
CRLB	Cramér-Rao lower bound
CT	Coordinated turn
CV	Constant velocity
EKF	Extended Kalman filter
FI	Fisher information
GLR	Generalized likelihood ratio
GPB	Generalized pseudo-Bayesian
HMM	Hidden Markov model
IA	Intrinsic accuracy
IEKF	Iterated extended Kalman filter
IID	Identically and independently distributed
IMM	Interacting multiple models
KF	Kalman filter
LSE	Least square estimate/estimator
MAP	Maximum <i>a priori</i>
MLE	Maximum likelihood estimate/estimator
MMSE	Minimum mean squares error
MSE	Mean square error
MVE	Minimum variance estimate/estimator
PDF	Probability density function
PF	Particle filter

---

Abbreviation	Meaning
RA	Relative accuracy
RMSE	Root mean square error
ROC	Receiver operating characteristics
SIR	Sampling importance resampling
SIS	Sequential importance sampling
SVD	Singular value decomposition
UKF	Unscented Kalman filter
UMP	Uniformly most powerful
UT	Unscented transform

## Symbols and Mathematical Notation

Notation	Meaning
$\mathbb{X}_t$	Stacked variables, $\mathbb{X}_t^T = (x_{t-L+1}^T, \dots, x_t^T)$ , if not stated otherwise, $L$ is chosen to include all available data. A similar notation is used for other stacked variables.
$x \sim y$	$x$ is distributed as $y$ .
$x \stackrel{a}{\sim} y$	$x$ is asymptotically distributed as $y$ .
$\nabla_x$	Gradient with respect to $x$ , see (A.1) below.
$\Delta_x^y$	Second derivative, $\nabla_y \nabla_x$ , see (A.2) below.
$L(\cdot) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma$	If $L(\cdot) \geq \gamma$ , reject $\mathcal{H}_0$ otherwise $\mathcal{H}_0$ is accepted.
$\Psi_x$	Relative accuracy of $x$ .
$\arg \max_x f(x)$	The $x$ maximizing $f(x)$ .
$\arg \min_x f(x)$	The $x$ minimizing $f(x)$ .
$\text{cov}(x)$	Covariance of $x$ .
$\delta(\cdot)$	Generalized Dirac function.
$\delta_t$	Mode indicator for time $t$ .
$\text{diag}(x_1, \dots, x_n)$	A diagonal matrix with $x_i$ in the diagonal.
$e_t$	Measurement noise at time $t$ .
$E(x)$	Expected value of $x$ .
$f(\cdot)$	State propagation function.
$f_t$	Deterministic but unknown fault at time $t$ .
$\mathcal{H}_i$	A hypothesis.
$h(\cdot)$	Measurement function.
$\mathcal{I}_x(\mu)$	Fisher information of $x$ with respect to $\mu$ .
$\mathcal{I}_x$	Intrinsic accuracy of $x$
$\mathcal{I}^{\text{KL}}(p(\cdot), q(\cdot))$	Kullback-Leibler information between the PDFs $p(\cdot)$ and $q(\cdot)$ .
$\mathcal{J}^{\text{K}}(p(\cdot), q(\cdot))$	Kullback divergence between the PDFs $p(\cdot)$ and $q(\cdot)$ .
$L(\cdot)$	Likelihood ratio.

Notation	Meaning
$L$	Window size.
$\mu_x$	Mean of $x$ .
$\mathcal{N}(x; \mu, \Sigma)$	Gaussian PDF for mean $\mu$ and covariance $\Sigma$ .
$\mathcal{N}_n(x; (\omega_\delta, \mu_\delta, \Sigma_\delta)_{\delta=1}^n)$	Gaussian mixture PDF with $n$ modes.
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean $\mu$ and covariance $\Sigma$ .
$\mathcal{N}_n((\omega_\delta, \mu_\delta, \Sigma_\delta)_{\delta=1}^n)$	Gaussian mixture distribution with $n$ modes.
$n_x$	Dimension of the variable $x$ .
$\omega_\delta$	Mode/particle probability.
$P_{t \tau}$	Estimation error covariance at time $t$ given the measurements $\mathbb{Y}_\tau$ .
$p(\cdot)$	Probability density function.
$\Pr(\mathcal{A})$	Probability of the statement $\mathcal{A}$ .
$Q_t$	Covariance of process noise at time $t$ .
$\mathcal{Q}_\star$	Complementary cumulative distribution function of the distribution $\star$ .
$R_t$	Covariance of measurement noise at time $t$ .
$r_t$	Residual
$\text{rank}(A)$	Rank of $A$ .
$\Sigma_x$	Covariance of $x$ .
$\text{tr}(A)$	Trace of $A$ .
$A^T$	$A$ transposed.
$u_t$	Known input at time $t$ .
$\text{var}(x)$	Variance of $x$ .
$w_t$	Process noise at time $t$ .
$x_t$	State at time $t$ .
$\hat{x}_{t \tau}$	Estimate of $x_t$ given the measurements $\mathbb{Y}_\tau$ .
$y_t$	Measurement at time $t$ .

## Definition of Derivatives

The derivative of  $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ , often denoted the *gradient* or the *Jacobian*, used is

$$\nabla_x f = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}. \quad (\text{A.1})$$

With this definition the second derivative of a scalar function  $f : \mathbb{R}^n \mapsto \mathbb{R}$  becomes

$$\Delta_x^y f = \nabla_x \nabla_y f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial y_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial y_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial y_m} \end{pmatrix}. \quad (\text{A.2})$$



---

## Bibliography

- [1] Daniel L. Alspach and Harold W. Sorenson. Recursive Bayesian estimation using Gaussian sum. *IEEE Transactions on Automatic Control*, 17(4):439–448, 1972.
- [2] Brian D. O. Anderson and John B. Moore. *Optimal Filtering*. Prentice-Hall, Inc, Englewood Cliffs, NJ., 1979. ISBN 0-13-638122-7.
- [3] Peter Andersson. Adaptive forgetting in recursive identification through multiple models. *International Journal of Control*, 42(5):1175–1193, 1985.
- [4] Christoph Arndt. *Information Measures*. Springer-Verlag, 2001. ISBN 3-540-40855-X.
- [5] Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc, 1993. ISBN 0-13-126780-9.
- [6] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *The Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. Reproduced in [10].
- [7] Niclas Bergman. *Bayesian Inference in Terrain Navigation*. Licentiate thesis no 649, Department of Electrical Engineering, Linköpings universitet, Sweden, 1997.
- [8] Niclas Bergman. *Recursive Bayesian Estimation: Navigation and Tracking Applications*. Dissertations no 579, Linköping Studies in Science and Technology, SE-581 83 Linköping, Sweden, May 1999.
- [9] Niclas Bergman, Lennart Ljung, and Fredrik Gustafsson. Terrain navigation using Bayesian statistics. *IEEE Control Systems Magazine*, 19(3):33–40, June 1999.
- [10] George A. Bernard and Thomas Bayes. Studies in the history of probability and statistics: IX. Thomas Bayes’s essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3/4):293–315, December 1958.

- [11] Henk A. P. Blom and Yaakov Bar-Shalom. The interacting multiple model algorithm for systems with Markovian switching coefficients. *IEEE Transactions on Automatic Control*, 33(8):780–783, August 1988.
- [12] Ben Zion Bobrovsky and Moshe Zakai. A lower bound on the estimation error for Markov processes. *IEEE Transactions on Automatic Control*, 20(6):785–788, December 1975.
- [13] Milodrag Bolić, Petar M. Djurić, and Sangjin Hong. Resampling algorithms and architectures for distributed particle filters. *IEEE Transactions on Signal Processing*, 53(7):2442–2450, July 2005.
- [14] Richard S. Bucy and Kenneth D. Senne. Digital synthesis of non-linear filters. *Automatica*, 7(3):287–298, 1971.
- [15] David R. Cox and David V. Hinkley. *Theoretical Statistics*. Chapman and Hall, New York, 1974.
- [16] Charlotte Dahlgren. Nonlinear black box modelling of JAS 39 Gripen’s radar altimeter. Master’s thesis no LiTH-ISY-EX-1958, Department of Electrical Engineering, Linköpings universitet, Sweden, October 1998.
- [17] Fred Daum and Jim Huang. Curse of dimensionality and particle filters. In *Proceedings of IEEE Aerospace Conference*, volume 4, pages 1979–1993, Big Sky, MT, USA, March 2003. IEEE.
- [18] Peter C. Doerschuck. Cramer-Rao bounds for discrete-time nonlinear filtering problems. *IEEE Transactions on Automatic Control*, 40(8):1465–1469, August 1995.
- [19] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10:197–208, 2000.
- [20] Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer-Verlag, New York, 2001. ISBN 0-387-95146-6.
- [21] Ronald A. Fisher. On the foundations of mathematical statistics. *The Philosophical Transactions of the Royal Society of London*, A(222):309–368, 1922.
- [22] Ronald A. Fisher. Theory of statistical estimation. In *Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725, 1925.
- [23] Philippe Forster and Pascal Larzabal. On the lower bounds for deterministic parameter estimation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1137–1140, Orlando, FL, USA, May 2002.
- [24] Jorge I. Galdos. A Cramér-Rao bound for multidimensional discrete-time dynamical systems. *IEEE Transactions on Automatic Control*, 25(1):117–119, February 1980.

- [25] Janos J. Gertler. *Fault Detection and Diagnosis in Engineering Systems*. Marcel Dekker, Inc, 1998. ISBN 0-8247-9427-3.
- [26] Gene H. Golub and Charles F. van Loan. *Matrix Computations*. John Hopkins University Press, 3 edition, 1996. ISBN 0-2018-54-14-8.
- [27] Neil J. Gordon, David J. Salmond, and Adrian F. M. Smith. Novel approach to non-linear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F Radar and Signal Processing*, 140(2):107–113, April 1993.
- [28] Fredrik Gustafsson. *Adaptive Filtering and Change Detection*. John Wiley & Sons, Ltd, Chichester, West Sussex, England, 2000. ISBN 0-471-49287-6.
- [29] Fredrik Gustafsson. Stochastic fault diagnosability in parity spaces. In *Proceedings of 15th Triennial IFAC World Congress on Automatic Control*, Barcelona, Spain, July 2002.
- [30] Fredrik Gustafsson, Fredrik Gunnarson, Niclas Bergman, Urban Forssell, Jonas Jansson, Rickard Karlsson, and Per-Johan Nordlund. Particle filters for positioning, navigation, and tracking. *IEEE Transactions on Signal Processing*, 50(2):425–437, February 2002.
- [31] Allan Gut. *An Intermediate Course in Probability*. Springer-Verlag, 1995. ISBN 0-387-94507-5.
- [32] John M. Hammersley and K. William Morton. Poor man’s Monte Carlo. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 16(1):23–38, 1954.
- [33] Gustaf Hendeby and Fredrik Gustafsson. On performance measures for approximate parameter estimation. In *Proceedings of Reglermöte 2004*, Chalmers, Gothenburgh, Sweden, May 2004.
- [34] Gustaf Hendeby and Fredrik Gustafsson. Fundamental filtering limitations in linear non-Gaussian systems. In *Proceedings of 16th Triennial IFAC World Congress*, Prague, Czech Republic, July 2005.
- [35] Gustaf Hendeby and Fredrik Gustafsson. Fundamental fault detection limitations in linear non-Gaussian systems. In *Proceedings of 44th IEEE Conference on Decision and Control and European Control Conference*, Sevilla, Spain, December 2005. To appear.
- [36] Jeroen D. Hol. Resampling in particle filters. Student thesis no. LiTH-ISY-EX-ET-0283-2004, Department of Electrical Engineering, Linköpings universitet, Sweden, May 2004.
- [37] Andrew H. Jazwinski. *Stochastic Processes and Filtering Theory*, volume 64 of *Mathematics in Science and Engineering*. Academic Press, Inc, 1970.
- [38] Simon J. Julier. The scaled unscented transformation. In *Proceedings of American Control Conference*, volume 6, pages 4555–4559, Anchorage, AK, USA, May 2002.

- [39] Simon J. Julier and Jeffery K. Uhlmann. Reduced sigma point filters for the propagation of means and covariances through nonlinear transformations. In *Proceedings of American Control Conference*, volume 2, pages 887–892, Anchorage, AK, USA, May 2002.
- [40] Simon J. Julier and Jeffrey K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, March 2004.
- [41] Simon J. Julier, Jeffrey K. Uhlmann, and Hugh F. Durrant-Whyte. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on Automatic Control*, 45(3), March 2000.
- [42] Thomas Kailath, Ali H. Sayed, and Babak Hassibi. *Linear Estimation*. Prentice-Hall, Inc, 2000. ISBN 0-13-022464-2.
- [43] Rudolph E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the American Society of Mechanical Engineering — Journal Basic Engineering*, 82(Series D):35–45, March 1960.
- [44] Rudolph E. Kalman and Richard S. Bucy. New results in linear filtering and prediction theory. *Transactions of the American Society of Mechanical Engineering — Journal Basic Engineering*, 83(Series D):95–108, March 1961.
- [45] Rickard Karlsson. *Particle Filtering for Positioning and Tracking Applications*. Dissertations no 924, Linköping Studies in Science and Technology, SE-581 83 Linköping, Sweden, March 2005.
- [46] Rickard Karlsson, Jonas Jansson, and Fredrik Gustafsson. Model-based statistical tracking and decision making for collision avoidance application. In *Proceedings of American Control Conference*, pages 3435–3440, Boston, MA, USA, July 2004.
- [47] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*, volume 1. Prentice-Hall, Inc, 1993. ISBN 0-13-042268-1.
- [48] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Detection Theory*, volume 2. Prentice-Hall, Inc, 1998. ISBN 0-13-504135-X.
- [49] Steven M. Kay and Dabasis Sengupta. Optimal detection in colored non-Gaussian noise with unknown parameter. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 1087–1089, Dallas, TX, USA, April 1987.
- [50] Augustine Kong, Jun S. Liu, and Wing H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, March 1994.
- [51] Jayesh H. Kotecha and Petar M Djurić. Gaussian particle filtering. *IEEE Transactions on Signal Processing*, 51(10):2592–2601, October 2003.
- [52] Stuart C. Kramer and Harold W. Sorenson. Bayesian parameter estimation. *IEEE Transactions on Automatic Control*, 33(2):217–222, February 1988.

- [53] Stuart C. Kramer and Harold W. Sorenson. Recursive Bayesian estimation using piece-wise constant approximations. *Automatica*, 24(6):789–801, November 1988.
- [54] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951.
- [55] Solomon Kullback, John C. Keegel, and Joseph H. Kullback. *Topics in Statistical Information Theory*, volume 42 of *Lecture Notes in Statistics*. Springer-Verlag, 1987. ISBN 0-387-96512-2.
- [56] Tine Lefebvre, Herman Bruyninckx, and Joris De Schutter. Comment on “A new method for the nonlinear transformation of means and covariances in filters and estimators”. *IEEE Transactions on Automatic Control*, 47(8), August 2002. Comments on [41].
- [57] Tine Lefebvre, Herman Bruyninckx, and Joris De Schutter. Kalman filters for nonlinear systems: a comparison of performance. *International Journal of Control*, 77(7), May 2004.
- [58] Eric L. Lehmann. *Theory of Point Estimation*. Probability and Mathematical Statistics. John Wiley & Sons, Ltd, 1983. ISBN 0-471-05849-1.
- [59] Eric L. Lehmann. *Testing Statistical Hypotheses*. Probability and Mathematical Statistics. John Wiley & Sons, Ltd, 2 edition, 1986. ISBN 0-471-84083-1.
- [60] Robert Leland. The Kulback-Leibler information divergence for continuous systems using white noise theory. In *Proceedings of 38th IEEE Conference on Decision and Control*, pages 1903–1907, Phoenix, AZ, USA, December 1999.
- [61] Rong X. Li and Vesslin P. Jilkov. Survey of maneuvering target tracking. part I: Dynamic models. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4): 1333–1364, October 2003.
- [62] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, January 1991.
- [63] Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, December 1971.
- [64] Gary Lorden. Open-ended tests for Koopman-Darmois families. *The Annals of Statistics*, 1(4):633–643, July 1973.
- [65] Jerzy Neyman and Egon S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20A(1/2):175–240, July 1928.
- [66] Jerzy Neyman and Egon S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika*, 20A(3/4):263–294, December 1928.

- [67] Man-Suk Oh. Monte Carlo integration via importance sampling: Dimensionality effect and an adaptive algorithm. In Nancy Flournoy and Robert K. Tsutakawa, editors, *Statistical Multiple Integration*, volume 115 of *Contemporary Mathematics*, pages 165–187. American Mathematical Society, Providence, RI, USA, 1991.
- [68] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill Book Co, 3 edition, 1991.
- [69] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286 1989.
- [70] C. Radhakrishna Rao. Minimum variance and the estimation of several parameters. *Proceedings of the Cambridge Philosophical Society*, 43:280–283, 1946.
- [71] Branko Ristic, Sanjeev Arulampalam, and Neil Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, Inc, 2004. ISBN 1-58053-631-X.
- [72] Christian P. Robert. *The Bayesian Choise: From Decision-Theoretic Foundation to Computational Implementation*. Springer texts in Statistics. Springer-Verlag, 2 edition, 2001.
- [73] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, 1999. ISBN 0-387-98707-X.
- [74] Wilson J. Rugh. *Linear System Theory*. Prentice-Hall, Inc, 1996. ISBN 0-13-441205-2.
- [75] Stanley F. Schmidt. Application of state-space methods to navigation problems. *Advances in Control Systems*, 3:293–340, 1966.
- [76] Debasis Sengupta and Steven M. Kay. Efficient estimation for non-Gaussian autoregressive processes. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(6):785–794, June 1989.
- [77] Gerald L. Smith, Stanley F. Schmidt, and Leonard A. McGee. Application of statistical filter theory to the optimal estimation of position and velocity on board a circumlunar vehicle. Technical Report NASA TR R-135, National Aeronautics and Space Administration, 1962.
- [78] Harold W. Sorenson, editor. *Kalman Filtering: Theory and Applications*. IEEE Press, 1985.
- [79] Harold W. Sorenson and Daniel L. Alspach. Recursive Bayesian estimation using Gaussian sums. *Automatica*, 7(4):465–479, July 1971.
- [80] Ondřej Straka and Miroslav Šimandl. Using the Bhattacharyya distance in functional sampling density of particle filter. In *Proceedings of 16th Triennial IFAC World Congress*, Prague, Czech Republic, July 2005.

- [81] Martin Svensson. Implementation and evaluation of bayesian terrain navigation. Master's thesis no LiTH-ISY-EX-2039, Department of Electrical Engineering, Linköpings universitet, Sweden, March 1999. In Swedish.
- [82] Peter Swerling. First order propagation in a stagewise smoothing procedure for satellite observations. *The Journal of the Astronautical Sciences*, 6:46–52, 1959.
- [83] James H. Taylor. The Cramér-Rao estimation lower bound computation for deterministic nonlinear systems. *IEEE Transactions on Automatic Control*, 24(2):343–344, April 1979.
- [84] Robert M. Taylor, Jr, Brian P. Flanagan, and John A. Uber. Computing the recursive posterior Cramer-Rao bound for a nonlinear nonstationary system. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume VI, pages 673–676, Hong Kong, April 2003. The conference was canceled.
- [85] Petr Tichavský, Carlos H. Muravchik, and Arye Nehorai. Posterior Cramér-Rao discrete-time nonlinear filtering. *IEEE Transactions on Signal Processing*, 46(5): 1386–1396, May 1998.
- [86] David Törnqvist, Fredrik Gustafsson, and Inger Klein. GLR tests for fault detection over sliding data windows. In *Proceedings of 16th Triennial IFAC World Congress*, Prague, Czech Republic, July 2005.
- [87] Rudolph van der Merwe and Eric A. Wan. The square-root unscented Kalman filter for state and parameter-estimation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3461–3464, Salt Lake City, UT, USA, May 2001.
- [88] Rudolph van der Merwe, Arnaud Doucet, Nando de Freitas, and Eric Wan. The unscented particle filter. Technical Report CUED/F-INFENG/TR 380, Cambridge Univeristy Engineering Department, August 2000.
- [89] Harry S. van Trees. *Part I. Detection, Estimation, and Modulation Theory*. Detection, Estimation, and Modulation Theory. John Wiley & Sons, Ltd, 1968. ISBN 0-471-89955-0.
- [90] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3):426–482, November 1943.
- [91] Eric A. Wan and Rudolph van der Merwe. The uncented Kalman filter for nonlinear estimation. In *Proceedings of IEEE Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pages 153–158, Lake Louise, AB, Canada, October 2000.
- [92] Yuanxin Wu, Dewen Hu, Meiping Wu, and Xiaoping Hu. Unscented Kalman filtering for additive noise case: Augmented versus nonaugmented. *IEEE Signal Processing Letters*, 12(5):357–360, May 2005.



**Licentiate Theses**  
**Division of Automatic Control**  
**Linköpings universitet**

- P. Andersson:** Adaptive Forgetting through Multiple Models and Adaptive Control of Car Dynamics. Thesis No. 15, 1983.
- B. Wahlberg:** On Model Simplification in System Identification. Thesis No. 47, 1985.
- A. Isaksson:** Identification of Time Varying Systems and Applications of System Identification to Signal Processing. Thesis No. 75, 1986.
- G. Malmberg:** A Study of Adaptive Control Missiles. Thesis No. 76, 1986.
- S. Gunnarsson:** On the Mean Square Error of Transfer Function Estimates with Applications to Control. Thesis No. 90, 1986.
- M. Viberg:** On the Adaptive Array Problem. Thesis No. 117, 1987.
- K. Ståhl:** On the Frequency Domain Analysis of Nonlinear Systems. Thesis No. 137, 1988.
- A. Skeppstedt:** Construction of Composite Models from Large Data-Sets. Thesis No. 149, 1988.
- P. A. J. Nagy:** MaMiS: A Programming Environment for Numeric/Symbolic Data Processing. Thesis No. 153, 1988.
- K. Forsman:** Applications of Constructive Algebra to Control Problems. Thesis No. 231, 1990.
- I. Klein:** Planning for a Class of Sequential Control Problems. Thesis No. 234, 1990.
- F. Gustafsson:** Optimal Segmentation of Linear Regression Parameters. Thesis No. 246, 1990.
- H. Hjalmarsson:** On Estimation of Model Quality in System Identification. Thesis No. 251, 1990.
- S. Andersson:** Sensor Array Processing; Application to Mobile Communication Systems and Dimension Reduction. Thesis No. 255, 1990.
- K. Wang Chen:** Observability and Invertibility of Nonlinear Systems: A Differential Algebraic Approach. Thesis No. 282, 1991.
- J. Sjöberg:** Regularization Issues in Neural Network Models of Dynamical Systems. Thesis No. 366, 1993.
- P. Pucar:** Segmentation of Laser Range Radar Images Using Hidden Markov Field Models. Thesis No. 403, 1993.
- H. Fortell:** Volterra and Algebraic Approaches to the Zero Dynamics. Thesis No. 438, 1994.
- T. McKelvey:** On State-Space Models in System Identification. Thesis No. 447, 1994.
- T. Andersson:** Concepts and Algorithms for Non-Linear System Identifiability. Thesis No. 448, 1994.
- P. Lindskog:** Algorithms and Tools for System Identification Using Prior Knowledge. Thesis No. 456, 1994.
- J. Plantin:** Algebraic Methods for Verification and Control of Discrete Event Dynamic Systems. Thesis No. 501, 1995.
- J. Gunnarsson:** On Modeling of Discrete Event Dynamic Systems, Using Symbolic Algebraic Methods. Thesis No. 502, 1995.
- A. Ericsson:** Fast Power Control to Counteract Rayleigh Fading in Cellular Radio Systems. Thesis No. 527, 1995.
- M. Jirstrand:** Algebraic Methods for Modeling and Design in Control. Thesis No. 540, 1996.
- K. Edström:** Simulation of Mode Switching Systems Using Switched Bond Graphs. Thesis No. 586, 1996.
- J. Palmqvist:** On Integrity Monitoring of Integrated Navigation Systems. Thesis No. 600, 1997.
- A. Stenman:** Just-in-Time Models with Applications to Dynamical Systems. Thesis No. 601, 1997.
- M. Andersson:** Experimental Design and Updating of Finite Element Models. Thesis No. 611, 1997.
- U. Forssell:** Properties and Usage of Closed-Loop Identification Methods. Thesis No. 641, 1997.

**M. Larsson:** On Modeling and Diagnosis of Discrete Event Dynamic systems. Thesis No. 648, 1997.

**N. Bergman:** Bayesian Inference in Terrain Navigation. Thesis No. 649, 1997.

**V. Einarsson:** On Verification of Switched Systems Using Abstractions. Thesis No. 705, 1998.

**J. Blom, F. Gunnarsson:** Power Control in Cellular Radio Systems. Thesis No. 706, 1998.

**P. Spångéus:** Hybrid Control using LP and LMI methods – Some Applications. Thesis No. 724, 1998.

**M. Norrlöf:** On Analysis and Implementation of Iterative Learning Control. Thesis No. 727, 1998.

**A. Hagenblad:** Aspects of the Identification of Wiener Models. Thesis No. 793, 1999.

**F. Tjärnström:** Quality Estimation of Approximate Models. Thesis No. 810, 2000.

**C. Carlsson:** Vehicle Size and Orientation Estimation Using Geometric Fitting. Thesis No. 840, 2000.

**J. Löfberg:** Linear Model Predictive Control: Stability and Robustness. Thesis No. 866, 2001.

**O. Härkegård:** Flight Control Design Using Backstepping. Thesis No. 875, 2001.

**J. Elbornsson:** Equalization of Distortion in A/D Converters. Thesis No. 883, 2001.

**J. Roll:** Robust Verification and Identification of Piecewise Affine Systems. Thesis No. 899, 2001.

**I. Lind:** Regressor Selection in System Identification using ANOVA. Thesis No. 921, 2001.

**R. Karlsson:** Simulation Based Methods for Target Tracking. Thesis No. 930, 2002.

**P.-J. Nordlund:** Sequential Monte Carlo Filters and Integrated Navigation. Thesis No. 945, 2002.

**M. Östring:** Identification, Diagnosis, and Control of a Flexible Robot Arm. Thesis No. 948, 2002.

**C. Olsson:** Active Engine Vibration Isolation using Feedback Control. Thesis No. 968, 2002.

**J. Jansson:** Tracking and Decision Making for Automotive Collision Avoidance. Thesis No. 965, 2002.

**N. Persson:** Event Based Sampling with Application to Spectral Estimation. Thesis No. 981, 2002.

**D. Lindgren:** Subspace Selection Techniques for Classification Problems. Thesis No. 995, 2002.

**E. Geijer Lundin:** Uplink Load in CDMA Cellular Systems. Thesis No. 1045, 2003.

**M. Enqvist:** Some Results on Linear Models of Nonlinear Systems. Thesis No. 1046, 2003.

**T. Schön:** On Computational Methods for Nonlinear Estimation. Thesis No. 1047, 2003.

**F. Gunnarsson:** On Modeling and Control of Network Queue Dynamics. Thesis No. 1048, 2003.

**S. Björklund:** A Survey and Comparison of Time-Delay Estimation Methods in Linear Systems. Thesis No. 1061, 2003.

**M. Gerdin:** Parameter Estimation in Linear Descriptor Systems. Thesis No. 1085, 2004.

**A. Eidehall:** An Automotive Lane Guidance System. Thesis No. 1122, 2004.

**J. Gillberg:** Methods for Frequency Domain Estimation of Continuous-Time Models. Thesis No. 1133, 2004.

**E. Wernholt:** On Multivariable and Nonlinear Identification of Industrial Robots. Thesis No. 1131, 2004.