Electrical Engineering and Automation

Iterative and Geometric Methods for State Estimation in Non-linear Models

---

Filip Tronarp

Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 12/2020

Iterative and Geometric Methods for State Estimation in Non-linear Models

Filip Tronarp

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall TU2 of the school on 31 January 2020 at 12:00.

Aalto University
School of Electrical Engineering
Electrical Engineering and Automation
Sensor Informatics and Medical Technology

Supervising professor
Professor Simo Särkkä, Aalto University, Finland

Thesis advisor
Professor Simo Särkkä, Aalto University, Finland

Preliminary examiner
Professor Lyudmila Mihaylova, University of Sheffield, United Kingdom and Professor Jimmy Olsson, Royal Institute of Technology, Sweden

Opponent
Professor Gustaf Hendeby, Linköping University, Sweden

Aalto University publication series
DOCTORAL DISSERTATIONS 12/2020

ISBN 978-952-60-8921-8 (printed)
ISBN 978-952-60-8922-5 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-8922-5

Unigrafia Oy
Helsinki 2020

Finland

Printed matter
4041-0619

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Filip Tronarp

Name of the doctoral dissertation
Iterative and Geometric Methods for State Estimation in Non-linear Models

Publisher School of Electrical Engineering

Unit Electrical Engineering and Automation

Series Aalto University publication series DOCTORAL DISSERTATIONS 12/2020

Field of research

Manuscript submitted 16 September 2019 **Date of the defence** 31 January 2020

Permission for public defence granted (date) 22 November 2019 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Many problems in science and engineering involve estimating a dynamic signal from indirect measurements subject to noise, where points can either evolve in continuous time or in discrete time. These problems are often formalised as inference in probabilistic state-space models, which are also frequently assumed to be Markovian. For inferring the value of the signal at a particular point in time, methods of inference can be divided into different classes, namely prediction, filtering (tracking), and smoothing. In prediction, only past measurements of the signal are used to infer its present value, whereas in filtering both past and present measurements are used, and in smoothing past, present, and future measurements are used. Prediction is useful in situations where decisions need to be made contingent on a future value of the signal before future measurements are made. On the other hand, filtering is useful when the signal needs to be inferred as the measurements arrive, that is on-line. Lastly, smoothing is the preferred choice when none of the aforementioned constraints are present as it allows the use of the entire sequence of measurements to infer the signal.

In this thesis, the filtering and smoothing problems and their applications are examined. In particular, iterative Gaussian filters and smoothers are developed for both inferring continuous and discrete time signals. Furthermore, it is shown that methods for inference in state-space models can be applied to the field of probabilistic numerics. More specifically, estimating the solutions to ordinary differential equations can be formulated as inference in a probabilistic state-space model, hence the solutions can be inferred using either Gaussian filtering methods or sequential Monte Carlo.

Another theme of this thesis is the exploitation of geometry - in a broad sense. Firstly, the geometry of probability densities, namely information geometry, is exploited to approximately infer the signal in filtering. Geometry is also exploited in terms of the geometry of the state-space, the space where the signal takes its values. That is, for tracking a time-varying unit vector, a continuous-time dynamic model is posed that respects the geometry of the unit sphere. Subsequently, a filtering algorithm is developed based on the von Mises--Fisher distribution for inference in this model. The method is demonstrated to have applications in tracking the local gravity and magnetic field vectors using a smartphone. Lastly, the geometry of Hilbert spaces is used to approximate a stochastic differential equation with an ordinary differential equation with random coefficients. On this basis filtering and smoothing algorithms are developed.

Keywords

ISBN (printed) 978-952-60-8921-8 **ISBN (pdf)** 978-952-60-8922-5

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2020

Pages 159 **urn** http://urn.fi/URN:ISBN:978-952-60-8922-5

Preface

After receiving the degree of Master of Science in Engineering, Engineering Mathematics in January 2016, I departed from my homeland of Sweden, a country sometimes best enjoyed at a distance, to begin the journey of writing this thesis under the excellent supervision of Prof. Simo Särkkä and the financial support of the Academy of Finland and Aalto ELEC Doctoral School. I am indebted to Prof. Erik Lindström with the Department of Mathematical Statistics at Lund University for finding me this opportunity and encouraging me to pursue it.

This work was conducted during my time as a doctoral candidate in the Sensor informatics and medical technology group at the Department of Electrical Engineering and Automation, Aalto University, in Finland. Though during my studies I have also enjoyed the brief but great hospitality of Prof. Ŏndrej Straka at University of West Bohemia in Plzeň, Prof. Tim Barfoot at Institute for Aerospace Studies, University of Toronto, Prof. Philipp Hennig at Max Planck Institute for Intelligent Systems, Tübingen, Prof. Chris Oates at the Alan Turing Institute, London, and Prof. Ángel García Fernández at the University of Liverpool.

During my time at Aalto I have met several people who in one way or another have aided in writing this thesis, either through fruitful discussions and collaborations or playful banter. I have shared office with Toni Karvonen, Marko Mikkonen, Kimmo Suotsalo, Jakub Prüher, and Gao Rui, the company of whom I have enjoyed in and out of office. In particular I have had the pleasure to discuss various technical topic with Toni Karvonen, which sometimes resulted in tangible results and at other times in the reminder of how awful maths can be. I also had the pleasure to collaborate with Jakub Prüher, though during my visit to the University of West Bohemia from whence he came. Although she has not convinced me of the benefits of hot water yet (without coffee beans!), my collaborations with Gao Rui has resulted in convincing applications of state estimation in optimisation.

I have enjoyed valuable discussions and collaboration with the former post-docs and group members Prof. Roland Hostettler and Prof. Ángel

García–Fernández, which have resulted in numerous publications, some of which are included in this thesis. I am thankful to the remaining group members Lauri Palva, Zenith Purisha, Zhao Zheng, Juha Sarmavuori, Ali Bahrami Rad, Muhammad Fuady Emzir, Joel Jaskari, and Matti Raitoharju who have contributed to both an enjoyable and stimulating atmosphere at work. I would also like to Prof. Lauri Parkkonen, Narayan Puthanmadam Subramaniyam, and Sara Sommariva at the Department of Neuroscience at Aalto University with whom I have had the pleasure to explore the fascinating difficulties of brain imaging.

I would also like to thank Marco Soldati for introducing me to bouldering and Marko Mikkonen, Dennis Yeung, Prof. Ivan Vujaklija, Silvia Cotreoni, Wendy Lam, and Liya Merzon for joining me with the recent hobby and obsession. An activity that has surely helped me keep sane and perhaps more importantly – fit.

Prof. Simo Särkkä whose hands-off approach to supervision has let me explore my own research interests rather freely, which I am grateful for. It has been encouraging and inspiring to have a supervisors whose enthusiasm is so readily obtained. I certainly could not have asked for a better supervisor.

Lastly, I would like to thank my family. I would not have gotten this far in life without their tremendous support. I look forward to continue sharing in their pains and joys for the rest of my life – because family is everything.

Helsinki, December 31, 2019,

Filip Tronarp

Contents

Preface 3

Contents 5

List of Publications 7

Author's Contribution 9

List of Figures 11

Abbreviations 13

Symbols 15

1. Introduction 17

2. Background 19
- 2.1 Statistical Inference 19
  - 2.1.1 Inference with a Gaussian Prior 21
  - 2.1.2 Statistical Linear Regression 23
  - 2.1.3 Moment Approximations 25
- 2.2 Information Geometry 27
  - 2.2.1 Statistical Manifolds 27
  - 2.2.2 $\mathscr{L}_2$-Projections onto the Tangent Space 27

3. Probabilistic State-Space Models 29
- 3.1 Continuous-Discrete-Time State-Space Models 30
  - 3.1.1 Stochastic Differential Equations 30
  - 3.1.2 Examples of Continuous-Discrete-Time State-Space Models 35
- 3.2 Discrete-Time State-Space Models 40
  - 3.2.1 Examples of Discrete-Time State-Space Models 41

4. Inference in Probabilistic State-Space Models 43

4.1 Bayesian Inference in Continuous-Discrete Models 44
    4.1.1 The Formal Solution 44
    4.1.2 Inference in Affine Models 46
    4.1.3 Approximate Inference in Non-Affine Models . . 47
4.2 Bayesian Inference in Discrete-Time Models 51
    4.2.1 The Formal Solution 52
    4.2.2 Inference in Affine Models 52
    4.2.3 Inference in Non-Affine Models 53

5. Discussion **63**
5.1 Publication I 63
    5.1.1 The Main Findings 64
    5.1.2 Reflections and Outlook 65
5.2 Publication II 66
    5.2.1 The Main Findings 66
    5.2.2 Reflections and Outlook 67
5.3 Publication III 68
    5.3.1 The Main Findings 69
    5.3.2 Reflections and Outlook 69
5.4 Publication IV 70
    5.4.1 The Main Findings 71
    5.4.2 Reflections and Outlook 71
5.5 Publication V 72
    5.5.1 The Main Findings 72
    5.5.2 Reflections and Outlook 73
5.6 Publication VI 74
    5.6.1 The Main Findings 75
    5.6.2 Reflections and Outlook 75

References **77**

Errata **85**

Publications **87**

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Filip Tronarp and Simo Särkkä. Iterative statistical linear regression for Gaussian smoothing in continuous-time non-linear stochastic dynamic systems. *Signal Processing*, Volume 159, pages 1–12, June 2019.

II Filip Tronarp, Ángel García–Fernández, and Simo Särkkä. Iterative Filtering and Smoothing in Nonlinear and Non-Gaussian Systems Using Conditional Moments. *IEEE Signal Processing Letters*, 25, 3, 408–412, March 2018.

III Filip Tronarp, Hans Kersting, Simo Särkkä, and Philipp Hennig. Probabilistic Solutions to Ordinary Differential Equations as Non-Linear Bayesian Filtering: A New Perspective. *Statistics and Computing*, Volume 29, Issue 6, pages 1297–1315, November 2019.

IV Filip Tronarp and Simo Särkkä. Updates in Bayesian Filtering by Continuous Projections on a Manifold of Densities. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, pages 5032–5036, May 2019.

V Filip Tronarp, Roland Hostettler, and Simo Särkkä. Continuous-Discrete von Mises-Fisher Filtering on S^2 for Reference Vector Tracking. In *Proceedings of the 21st International Conference on Information Fusion (FUSION)*, Cambridge, UK, pages 1345–1352, July 2018.

VI Filip Tronarp and Simo Särkkä. Non-Linear Continuous-Discrete Smoothing By Basis Function Expansions Of Brownian Motion. In *Proceedings of the 21st International Conference on Information Fusion*

(FUSION), Cambridge, UK, pages 1114–1121, July 2018.

Author's Contribution

Publication I: "Iterative statistical linear regression for Gaussian smoothing in continuous-time non-linear stochastic dynamic systems"

The original idea is due to Simo Särkkä, which was expanded by Filip Tronarp. Filip Tronarp wrote the article and produced the code for the experiments with valuable feedback from Simo Särkkä.

Publication II: "Iterative Filtering and Smoothing in Nonlinear and Non-Gaussian Systems Using Conditional Moments"

The original idea is due to Filip Tronarp. Filip Tronarp also wrote the article and produced the code for the experiments with valuable feedback from Ángel García–Fernández and Simo Särkkä.

Publication III: "Probabilistic Solutions to Ordinary Differential Equations as Non-Linear Bayesian Filtering: A New Perspective"

The measurement model formulation is due to Filip Tronarp, though Simo Särkkä has tried to argue along similar lines previously. Filip Tronarp wrote most of the article with great help by Hans Kersting in fleshing out the introduction and the discussion. Filip Tronarp produced the code for the experiments. Simo Särkkä and Philipp Henning provided valuable feedback along the way.

Publication IV: "Updates in Bayesian Filtering by Continuous Projections on a Manifold of Densities"

Filip Tronarp and Simo Särkkä came up with the original idea independently from one another. Filip Tronarp wrote the article and produced the code for the experiments with valuable feedback from Simo Särkkä.

Publication V: "Continuous-Discrete von Mises-Fisher Filtering on S^2 for Reference Vector Tracking"

The original idea is due to Filip Tronarp who wrote most of the article and produced the code for the simulation experiments. Roland Hostettler designed and implemented the smartphone experiments and wrote the corresponding sections of the article. Simo Särkkä gave valuable feedback along the way.

Publication VI: "Non-Linear Continuous-Discrete Smoothing By Basis Function Expansions Of Brownian Motion"

Simo Särkkä came up with the original idea Filip Tronarp wrote the article and produced the code for the experiments with valuable feedback from Simo Särkkä.

Language check

The language of my dissertation has been checked by Doctoral candidates Dennis Yeung and Marco Soldati. I have personally examined and accepted/rejected the results of the language check one by one. This has not affected the scientific content of my dissertation.

List of Figures

1.1 Example of a signal in noise. 18

2.1 An illustration of Bayes' rule. The prior (dashed line) is weighted by the likelihood (dotted line) to form the posterior (solid line). 20

3.1 100 independent realisations of a Wiener process. 31
3.2 The convergence of $\hat{W}_L$ to W for $L = 1,\dots,2^8$. $\hat{W}_L$ is shown to the left and the root mean square error (RMSE) is shown to the right. 35
3.3 One realisation of the q times integrated Wiener process ($X^{(1)}$) on $\mathbb{R}$ for $q = 0,1,2,3$. 37
3.4 Example realisation of the radar tracked coordinated turn model. 39
3.5 One realisation of the stochastic rotation model in Equation (3.20) with $\breve{\Omega} = 0$ (Spherical Wiener process). 40
3.6 One realisation of the stochastic volatility model. The signal is on the left and the measurements are on the right. 42

Abbreviations

ADF	Assumed density filter
ADS	Assumed density smoother
ATI	Affine time-invariant
ATV	Affine time-variant
KF	Kalman filter
LTI	Linear time-invariant
LTV	Linear time-variant
ODE	Ordinary differential equation
pdf	Probability density function
PF	Particle filter
RMSE	Root mean-square error
SDE	Stochastic differential equation
SMC	Sequential Monte Carlo

Symbols

$\mathbb{C}[\cdot,\cdot]$	Cross-covariance operator
diag	The (block-)matrix formed by placing the arguments on the diagonal
$\mathbb{E}[\cdot]$	Expectation operator
J_c	Jacobian of the function c
$\mathscr{L}_2(\mathbb{X})$	The space of square integrable functions with domain $\mathbb{X}$
$\mathbb{N}$	Natural numbers
$\mathbb{N}_0$	Natural numbers (including zero)
p_X	Marginal probability density of X
$p_{Y,X}$	Joint probability density of X and Y
$p_{X \mid Y}$	Conditional probability density of X given the outcome y of Y
$\mathbb{R}$	Real numbers
$\mathbb{R}_+$	Positive real numbers (including zero)
span	The vector space spanned by the arguments
$\mathrm{tr}[\cdot]$	Trace operator
$\mathbb{V}[\cdot]$	Covariance matrix operator
$\delta(x-x_0)$	The dirac delta function at x_0
$\dot{\phi}$	Partial derivative of ϕ with respect to time (t), $\dot{\phi}(t)=\partial_t\phi(t)$
∂	Partial derivative
∇	Vector of partial differentials
$\times$	Cross-product or Cartesian product
$[\cdot]_\times$	Cross-product matrix, $[u]_\times v = u \times v$
$\otimes$	Kronecker's product
$\sim$	Distributed as, for example, $X \sim \mathcal{N}(\mu,\Sigma)$
$\dot{\sim}$	Approximately distributed as
$\wedge$	For $x,y \in \mathbb{R}$, $x \wedge y$ is the smallest number of x and y
$\langle\cdot,\cdot\rangle$	Inner product
$\circ$	Composition operator for two maps

1. Introduction

A plethora of problems in science and engineering can be formalised as a “signal in noise” problem, see Figure 1.1 for an example. That is, there is an *unknown signal*, which is measured indirectly by some device. The goal is to reconstruct the signal, however the measurement does not provide complete information about the signal and tends to be contaminated with “noise”. Therefore, from a science and engineering point of view, a principled way of inferring the signal is required, which is the formal term for hazarding a guess. In the context of statistical signal processing, this means that the properties of the signal and measurements are characterised probabilistically, that is to say that the signal and measurements are assumed to be outcomes of a probability model. Once an appropriate model to characterise the system has been identified, inferring the signal is formally solved by the application of Bayes’ rule.

This thesis is focused on inference in *stochastic dynamic systems*. That is, a signal is a quantity that varies over time and is measured periodically. This type of problem occurs frequently in many domains, such as tracking, navigation, processing of audio signals, and finance (Bar-Shalom et al., 2004, Godsill and Rayner, 1998, Lindström et al., 2015, Stone et al., 2014, Titterton and Weston, 2004). More specifically, the probability models considered in this thesis are of so-called state-space type. Inference in state-space models is a well studied topic and several books have been written over the past fifty years, for example, Anderson and Moore (1979), Bar-Shalom et al. (2004), Cappé et al. (2005), Crassidis and Junkins (2004), Gelb (1974), Jazwinski (1970), Maybeck (1979,1982,1982), Särkkä (2013), Särkkä and Solin (2019), Simon (2006), which ought not to be taken for an exhaustive list.

This thesis comprises of Publications I through VI and this introduction to the research field. Detailed accounts of this thesis’ contributions are given in the original publications. The purpose of this introduction is to give a brief description of the research field and establish the milieu in which the contributions lie. This serves as a basis for discussing the significance of the contributions and their relationship with the present

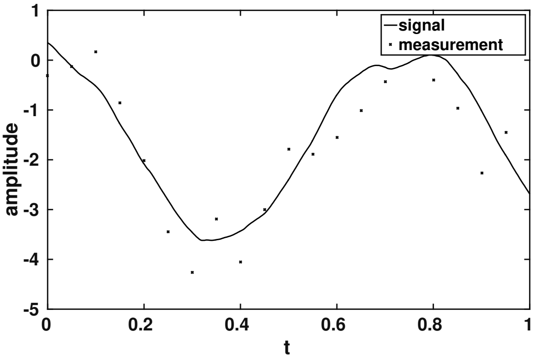

Figure 1.1. Example of a signal in noise.

scientific literature as well as pointing out future research directions.

The rest of this thesis is organised as follows. In Chapter 2, the basic statistical toolkit used in state estimation is reviewed along with the more esoteric concept of information geometry. Chapter 3 goes through the dynamic signal and measurement models that this thesis is concerned with, the combination of which define probabilistic state-space models. In Chapter 4 the problem of inference in state-space models is discussed. The formal filtering and smoothing relations are given as well as the special cases that can be solved by Kalman filtering and Rauch–Tung–Striebel smoothing. Furthermore, an account of approximate inference methods that are related to this thesis is also given. In Chapter 5, the key contributions of each publication are discussed, their relation with present scientific literature, and some future research directions are pointed out.

2. Background

In this chapter a description of basic concepts in statistical inference as it pertains to signal processing is given. The basic building blocks are the *prior probability density*, the measurement likelihood, and Bayes' rule, the latter determines how the former two ought to be combined to form an estimate or *posterior probability distribution*. Some common approximations to Bayes' rule when the prior is Gaussian are also reviewed. Furthermore, the slighty more esoteric subject of field of information geometry is reviewed in brief.

2.1 Statistical Inference

While the primary concern of this thesis is that of state estimation, the case of estimation in static systems is an important subproblem. Indeed, the so-called filter update in continuous-discrete and discrete-time state estimation is such a problem. Therefore, this section is dedicated to the case of static estimation.

The static estimation problem involves two random variables X and Y taking values in $\mathbb{X} \subset \mathbb{R}^d$ and $\mathbb{Y} \subset \mathbb{R}^m$, respectively, with joint probability density $p_{X,Y}$ and the inference problem is then to infer X given that the outcome of Y is y. In the Bayesian sense this entails computing the probability density of X conditioned on the outcome of Y being y, which is defined as

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

where p_Y is the marginal probability density of Y, which is given by

$$p_Y(x) = \int_{\mathbb{X}} p_{X,Y}(x,y) \, \mathrm{d}x.$$

The conditional density $p_{X|Y}$ can be expressed in terms of p_X and $p_{Y|X}$, which is Bayes' rule given in Theorem 1.

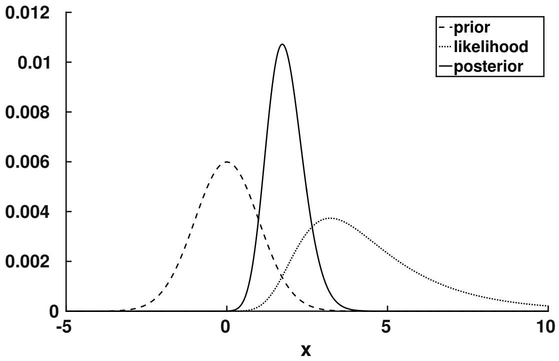

Figure 2.1. An illustration of Bayes' rule. The prior (dashed line) is weighted by the likelihood (dotted line) to form the posterior (solid line).

Theorem 1 (Bayes' Rule)**.** *Let* $X \sim p_X(x)$ *and* $Y \mid X \sim p_{Y|X}(y \mid x)$*. Then the probability density of* $X \mid Y$ *is given by*

$$p_{X|Y}(x \mid y) = \frac{p_{Y|X}(y \mid x) p_X(x)}{\int_{\mathbb{X}} p_{Y|X}(y \mid x) p_X(x) \, \mathrm{d}x} = \frac{p_{Y|X}(y \mid x) p_X(x)}{p_Y(y)}.$$

Due to Bayes' rule and other modelling related conveniences, the model $p_{X,Y}$ is typically formulated in terms of p_X and $p_{Y|X}$ directly

$$X \sim p_X(x),$$
$$Y \mid X \sim p_{Y|X}(x).$$

In the language of Bayesian statistics, the density $p_X(x)$ is the prior density for X and the conditional density $p_{X|Y}(x \mid y)$ is referred to as the posterior density for X given the outcome y of Y. Furthermore, for each fixed y the quantity $p_{Y|X}(y \mid x)$ is a function of x, which is referred to as the likelihood function. Throughout this thesis the likelihood function and its logarithm will frequently be referred to as $L(x)$ and $\ell(x)$, respectively,

$$L(x) \triangleq p_{Y|X}(y \mid x),$$
$$\ell(x) \triangleq \log p_{Y|X}(y \mid x).$$

A graphic illustration of Bayes' rule is shown in Figure. 2.1.

Point Estimation
The conditional density $p_{X|Y}$ gives a complete description of the probabilistic properties of X given the outcome y of Y. However, it is often the case that a single quantity is sought to represent an estimate as to what the outcome of X is. A survey on the different approaches to doing this is not given here, but rather the two most common options are given in the following.

Conditional mean estimate. The conditional mean estimate, as the name suggests estimates X by its conditional mean, which is given by

$$\mathbb{E}[X \mid Y = y] = \int_{\mathbb{X}} x p_{X|Y}(x \mid y) \mathrm{d}x.$$

The conditional mean is the classical Bayes' estimator under mean square error risk (Lehmann and Casella, 2006). Furthermore, the conditional variance defines an assessment of uncertainty in this estimate, which is given by

$$\mathbb{V}[X \mid Y = y] = \int_{\mathbb{X}} x x^{\mathsf{T}} p_{X|Y}(x \mid y) \mathrm{d}x - \mathbb{E}[X \mid Y]\mathbb{E}[X \mid Y]^{\mathsf{T}}$$

Maximum a posteriori estimate. The maximum a posteriori estimate of X is defined as the x that maximises $p_{X|Y}$. That is,

$$\begin{aligned}\hat{x}_{\mathrm{MAP}} &\triangleq \arg\max_{x \in \mathbb{X}} p_{X|Y}(x \mid y) \\ &= \arg\max_{x \in \mathbb{X}} \big(\ell(x) + \log p_X(x)\big).\end{aligned}$$

Remark 1. *Unless the plausibility of misunderstandings is fair, the outcome y is in this thesis omitted from the notation of conditional moments. For example, $\mathbb{E}[X \mid Y = y] = \mathbb{E}[X \mid Y]$ and $\mathbb{V}[X \mid Y = y] = \mathbb{V}[X \mid Y]$.*

2.1.1 Inference with a Gaussian Prior

Practitioners of stochastic signal processing have a conspicuous habit of getting themselves into inferential problems where their prior for some quantity X is Gaussian. That is, the probability model is given by

$$X \sim \mathcal{N}(\mu, \Sigma), \tag{2.3a}$$

$$Y \mid X \sim p_{Y|X}(y \mid x). \tag{2.3b}$$

The types of models of the form in Equation (2.3) can be divided into three relevant and overlapping classes, which are listed in the following.

Affine Gaussian System. In an affine Gaussian system, the measurement model in Equation (2.3) is given by

$$Y \mid X \sim \mathcal{N}\big(CX + d, R\big). \tag{2.4}$$

Conditionally Gaussian System. In a conditionally Gaussian system, the measurement model in Equation (2.3) is given by

$$Y \mid X \sim \mathcal{N}\big(c(X), R(X)\big).$$

Third Class. The third class, covers all the measurement models that are not conditionally Gaussian.

Approaches to Gaussian inference of the first two classes are reviewed in the following. The third class is too broad to give a comprehensive and unified treatment of, though some approaches for this scenario are presented in Publication II and Publication IV.

Affine Gaussian Systems

The case of affine Gaussian systems is particularly fortunate since the posterior $p_{X|Y}$ remains in the class of Gaussian distributions. The particular form of this relationship is given in Lemma 1 (see, e.g., Särkkä 2013, Lemma A.1 and A.2).

Lemma 1. *Let* X,Y *be random variables governed by the model in Equation* (2.3)*, with the measurement model given by Equation* (2.4) $c(X) = CX + d$*, and* $R(X) = R$*. Then* X *and* Y *are jointly Gaussian and* $X \mid Y \sim \mathcal{N}(\mu^+, \Sigma^+)$*, where*

$$S = C\Sigma C^\top + R, \tag{2.5a}$$

$$K = \Sigma C^\top S^{-1}, \tag{2.5b}$$

$$\mu^+ = \mu + K\big(y - C\mu - d\big), \tag{2.5c}$$

$$\Sigma^+ = \Sigma - KSK^\top. \tag{2.5d}$$

When the prior is in some class of probability distributions and the likelihood is such that the posterior remains in this class, the class is said to be *conjugate* with respect to the likelihood.

Conditionally Gaussian Measurements

Inference is in general intractable for conditionally Gaussian systems. However, by writing Equation (2.5) in terms of the joint moments of X and Y,

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X] + \mathbb{C}[X,Y]\mathbb{V}[Y]^{-1}\big(y - \mathbb{E}[Y]\big), \tag{2.6a}$$

$$\mathbb{V}[X \mid Y] = \mathbb{V}[X] - \mathbb{C}[X,Y]\mathbb{V}[Y]^{-1}\mathbb{C}[X,Y]^\top, \tag{2.6b}$$

an approximate approach is made apparent. That is, the joint moments of X and Y are computed and inserted into the right-hand side of Equation (2.6), which gives the following approximation.

Approximation 1 (Moment matching update)**.** *The approximation is* *given by* $X \mid Y \mathrel{\dot\sim} \mathcal{N}(\mu^+, \Sigma^+)$, *where*

$$S = \mathbb{V}[c(X)] + \mathbb{E}[R(X)], \tag{2.7a}$$

$$K = \mathbb{C}[X, c(X)]S^{-1}, \tag{2.7b}$$

$$\mathbb{E}[X \mid Y] \approx \mu^+ = \mu + K\big(y - \mathbb{E}[c(X)]\big), \tag{2.7c}$$

$$\mathbb{V}[X \mid Y] \approx \Sigma^+ = \Sigma - KSK^\top, \tag{2.7d}$$

where $S = \mathbb{V}[Y]$ *by the law of total variance.*

Approximation 1 is known as Gaussian *moment matching*, which means that a Gaussian distribution is matched to the moments of (X, Y) (Särkkä, 2013, Chapter 6).

2.1.2 Statistical Linear Regression

An important concept is that of statistical linear regression, which in the context of state estimation was introduced by Lefebvre et al. (2002) (statistical linearisation is a precursor concept, see Gelb 1974). It is a different way of deriving Approximation 1, but it has further implications as well. Namely, it can be used to define an iterative update, which has come to be known as posterior linearisation (García-Fernández et al., 2014, 2015).

Suppose that the random variables X and Y are taking values in $\mathbb{R}^d$ and $\mathbb{Y} \subset \mathbb{R}^m$, respectively, and are governed by the following probability model

$$X \sim p_X(x),$$

$$Y \mid X \sim p_{Y\mid X}(y \mid x).$$

The purpose of statistical linear regression is to find an affine representation of Y in terms of X. That is,

$$Y = CX + d + E,$$

for some matrix $C \in \mathbb{R}^{k\times d}$, vector $d \in \mathbb{R}^k$, and random variable $E \in \mathbb{R}^k$. While there are of course several such affine representations, interest lies in those that are optimal in some sense. In particular, the mean square optimal representation is found by minimising $\mathbb{E}\big[\lVert E\rVert^2\big]$ with respect to C and d. The resulting parameters are then given by

$$\hat{C} = \mathbb{C}[Y, X]\mathbb{V}[X]^{-1},$$

$$\hat{d} = \mathbb{E}[Y] - \hat{C}\mathbb{E}[X],$$

and the residual $E = Y - \hat{C}X - \hat{d}$ becomes a zero mean random variable with covariance matrix given by

$$\hat{R} \triangleq \mathbb{V}[E] = \mathbb{V}[Y] - \hat{C}\mathbb{V}[X]\hat{C}^\top.$$

Statistical Linear Regression in Conditionally Gaussian Systems

If $p_X(x) = \mathcal{N}(x;\mu,\Sigma)$ and $p_{Y|X}(y \mid x) = \mathcal{N}(y;c(x),R(x))$ then statistical linear regression gives a different derivation of Approximation 1. In this case, the parameters are given by

$$\begin{aligned}
\hat{C} &= \mathbb{C}[c(X),X]\Sigma^{-1},\\
\hat{d} &= \mathbb{E}[c(X)] - \hat{C}\mu,\\
\hat{R} &= \mathbb{V}[c(X)] + \mathbb{E}[R(X)] - \hat{C}\Sigma\hat{C}^{\top},
\end{aligned}$$

which gives the following method for approximate inference in the conditionally Gaussian system.

Approximation 2 (Statistical linear regression update)**.** *The approximation is given by* $X \mid Y \mathrel{\dot\sim} \mathcal{N}(\mu^+,\Sigma^+)$*, where*

$$\begin{aligned}
S &= \hat{C}\Sigma\hat{C}^{\top} + \hat{R},\\
K &= \Sigma\hat{C}^{\top}S^{-1},\\
\mathbb{E}[X \mid Y] \approx \mu^+ &= \mu + K\big(y - \hat{C}\mu - \hat{d}\big),\\
\mathbb{V}[X \mid Y] \approx \Sigma^+ &= \Sigma - KSK^{\top}.
\end{aligned}$$

Remark 2. *It is easy to verify that Approximation 2 and Approximation 1 are indeed the same.*

Posterior Linearisation

In essence, statistical linear regression uses the density p_X to linearise the relationship between X and Y that is implied by $p_{Y|X}$. This means that p_X can be thought of as a linearisation point, which is the important insight behind the posterior linearisation update (García-Fernández et al., 2014, 2015). The trick is to alternate between approximate updating according to Approximation 2 and updating the linearisation point, which results in Approximation 3.

Approximation 3 (Iterated statistical linear regression update)**.** *The approximation is given by* $X \mid Y \mathrel{\dot\sim} \mathcal{N}(\mu^+,\Sigma^+)$*, where* (μ^+,Σ^+) *are defined as the fixed-point to the following iteration.*

$$\begin{aligned}
S^l &= \hat{C}^l\Sigma\big[\hat{C}^l\big]^{\top} + \hat{R}^l,\\
K^l &= \Sigma\big[\hat{C}^l\big]^{\top}\big[S^l\big]^{-1},\\
\mu^{l+1} &= \mu + K^l\big(y - \hat{C}^l\mu - \hat{d}^l\big),\\
\Sigma^{l+1} &= \Sigma - K^lS^l\big[K^l\big]^{\top}.
\end{aligned}$$

where

$$\hat{C}^l = \mathbb{C}^l[c(X), X]\left[\Sigma^l\right]^{-1}, \tag{2.13a}$$

$$\hat{d}^l = \mathbb{E}^l[c(X)] - \hat{C}^l \mu^l, \tag{2.13b}$$

$$\hat{R}^l = \mathbb{V}^l[c(X)] + \mathbb{E}^l[R(X)] - \hat{C}^l \Sigma \left[\hat{C}^l\right]^\top, \tag{2.13c}$$

and $\mathbb{E}^l$, $\mathbb{C}^l$, *and* $\mathbb{V}^l$ *are with respect to* $X \sim \mathcal{N}(\mu^l, \Sigma^l)$. *Finally, the iterations are initialised at the prior* $(\mu^0, \Sigma^0) = (\mu, \Sigma)$.

Remark 3. *Originally, Approximation 3 was only derived for the case when* $R(x)$ *is constant,* $R(x) = R$ *(García-Fernández et al., 2014, 2015). The extension to non-constant* $R(x)$ *is in fact part of the contribution of Publication II.*

Remark 4. *If* $R(x) = R$ *and the expectations in Equation* (2.13) *are approximated by the Taylor series method (see Section 2.1.3) then Approximation 3 is the Gauss–Newton method for the maximum a posteriori estimate (Bell and Cathey, 1993).*

2.1.3 Moment Approximations

The methods discussed previously in Sections 2.1.1 and 2.1.2, and indeed the methods that will be discussed in Chapter 4 require computing expectations of various non-linear functions of $X \sim \mathcal{N}(\mu, \Sigma)$. This is often intractable, which is why approximations are required. The two most popular classes of moment approximation methods are the Taylor series methods and the cubature methods. Whenever a potentially intractable expectation appears in the sequel and indeed the prequel, it should be understood that it is approximated by any of these aforementioned methods, which are described below.

The Taylor Series Approach
The Taylor series approach, as the name suggests, involves expanding c and R in Taylor series. This is typically done up to first order around the prior mean μ, which gives

$$c(X) \approx c(\mu) + J_c(\mu)(X - \mu),$$

$$R(X) \approx R(\mu) + \sum_{i=1}^{d} \partial_i R(\mu)(X_i - \mu_i).$$

This results in the following approximation to the moments in Equation (2.7).

Approximation 4 (First order Taylor series approximation)**.**

$$\begin{aligned}\mathbb{E}[Y] &\approx c(\mu),\\ \mathbb{C}[X,Y] &\approx \Sigma J_c^{\top}(\mu),\\ \mathbb{V}[c(X)] &\approx J_c(\mu)\Sigma J_c^{\top}(\mu),\\ \mathbb{E}[R(X)] &\approx R(\mu).\end{aligned}$$

This approximation can of course be refined by including higher order terms in the Taylor series (see, e.g., Särkkä 2013, Section 5.2 and Gustafsson and Hendeby 2012).

The Cubature Approach

Numerical integration or cubature involves approximating expectations with respect to $X \sim \mathcal{N}(\mu,\Sigma)$ by selecting a set of nodes $\{\mathcal{X}_l\}_{l=1}^{L}$ with associated weights $\{w_l\}_{l=1}^{L}$. Then the expectation of some function $\phi(X)$ is approximated via the cubature rule

$$\mathbb{E}[\phi(X)] \approx \sum_{l=1}^{L} w_l \phi(\mathcal{X}_l),$$

which results in the following approximation.

Approximation 5 (Cubature approximation)**.**

$$\begin{aligned}\mathbb{E}[c(X)] &\approx \sum_{l=1}^{L} w_l c(\mathcal{X}_l),\\ \mathbb{C}[X,c(X)] &\approx \sum_{l=1}^{L} w_l (\mathcal{X}_l - \mu) c^{\top}(\mathcal{X}_l),,\\ \mathbb{V}[c(X)] &\approx \sum_{l=1}^{L} w_l c(\mathcal{X}_l) c^{\top}(\mathcal{X}_l) - \left(\sum_{l=1}^{L} w_l c(\mathcal{X}_l)\right)\left(\sum_{l=1}^{L} w_l c(\mathcal{X}_l)\right)^{\top},\\ \mathbb{E}[R(X)] &\approx \sum_{l=1}^{L} w_l R(\mathcal{X}_l).\end{aligned}$$

The nodes and weights of the cubature rule are often selected such that expectations of polynomials up to some order are computed exactly, such as in tensor products of Gauss–Hermite quadratures (Golub and Welsch, 1969) or fully symmetric cubature (McNamee and Stenger, 1967). The third degree rule of the latter came to be known as the unscented transform when it was introduced to the signal processing community (Julier and Uhlmann, 2004, Julier et al., 1995). Recently a class of cubature methods known as *Bayesian cubature* or *Gaussian process cubature* has started to gained traction in the signal processing community as well (Karvonen and Särkkä, 2017, Prüher and Straka, 2017, Prüher and Šimandl, 2015, Särkkä et al., 2015).

2.2 Information Geometry

Information geometry is the study of probability models using differential geometry (Amari, 2012, Amari and Nagaoka, 2007). While the field contains a fairly vast amount of material, the following presentation only includes the bare minimum for the purposes of this thesis. The interest lies in the projection methods of Brigo et al. (1999) (see also Brigo et al. 1998, Koyama 2018). That is, the following account is based on Brigo et al. (1999), though excluding a lot of the rigour.

2.2.1 Statistical Manifolds

Consider the family of probability densities on $\mathbb{R}^d$

$$\mathscr{P} = \left\{ p_\theta \colon \theta \in \Theta \subseteq \mathbb{R}^\nu \right\}.$$

A family of square root densities associated with $\mathscr{P}$ is defined as

$$\mathscr{P}^{1/2} = \left\{ p_\theta^{1/2} \colon p_\theta \in \mathscr{P} \right\}$$

and a homeomorphism $\phi \colon \mathscr{P}^{1/2} \mapsto \Theta$ with inverse $\phi^{-1}(\theta) = p_\theta^{1/2}$. In the language of differential geometry the pair $(\mathscr{P}^{1/2}, \phi)$ forms a chart of some ν-dimensional manifold $\mathscr{Q} \subset \mathscr{R} \subset \mathscr{L}_2(\mathbb{R}^d)$, where

$$\mathscr{R} = \left\{ p^{1/2} \colon p^{1/2} \in \mathscr{L}_2(\mathbb{R}^d),\ p(x) \geq 0 \right\}.$$

For present purposes, a single chart is sufficient. Furthermore, the tangent space to $\mathscr{P}^{1/2}$ at $p_\theta^{1/2}$ is given by the set of m linearly independent vectors

$$L_{p_\theta^{1/2}} \mathscr{P}^{1/2} = \operatorname{span} \left\{ \partial_1 p_\theta^{1/2}, \ldots, \partial_\nu p_\theta^{1/2} \right\} \subseteq \mathscr{L}_2(\mathbb{R}^d),$$

where the partial derivatives are with respect to θ. The tangent space $L_{p_\theta^{1/2}} \mathscr{P}^{1/2}$ inherits the inner product of $\mathscr{L}_2(\mathbb{R}^d)$ and for the basis elements of $L_{p_\theta^{1/2}} \mathscr{P}^{1/2}$ it holds that

$$\left\langle \partial_i p_\theta^{1/2}, \partial_j p_\theta^{1/2} \right\rangle = \frac{1}{4} g_{ij}(\theta),$$

where $g(\theta)$ is the Fisher information matrix.

2.2.2 $\mathscr{L}_2$-Projections onto the Tangent Space

With the minimal setup of the preceding section, the orthogonal projection of $\mathscr{L}_2(\mathbb{R}^d)$ onto $L_{p_\theta^{1/2}} \mathscr{P}^{1/2}$ can be defined. Since the chosen basis for $L_{p_\theta^{1/2}} \mathscr{P}^{1/2}$ is not orthonormal, the projection operator $\Pi_\theta \colon \mathscr{L}_2(\mathbb{R}^d) \mapsto L_{p_\theta^{1/2}} \mathscr{P}^{1/2}$ takes the following form

$$\Pi_\theta v = \sum_{i,j} 4 g_{ij}^{-1}(\theta) \left\langle v, \partial_j p_\theta^{1/2} \right\rangle \partial_i p_\theta^{1/2}.$$

If the element v being projected has a particular form then the projection formula simplifies according to the following lemma from Brigo et al. (1999).

Lemma 2. *Let the function u satisfy*

$$\mathbb{E}_\theta\left[|u|^2\right] < \infty,$$

where $\mathbb{E}_\theta$ denotes the expectation with respect to p_θ. Then $v = \frac{1}{2} p_\theta^{1/2} u \in \mathscr{L}_2(\mathbb{R}^d)$ and the projection formula takes the form

$$\Pi_\theta v = \sum_{i,j} g_{ij}^{-1}(\theta) \mathbb{E}_\theta\left[u \partial_j \log p_\theta\right] \partial_i p_\theta^{1/2}.$$

It is these projection formulae which form the basis of the projection filters and smoothers (Brigo et al., 1998, 1999, Koyama, 2018), the principles of which are exploited in Publication IV.

3. Probabilistic State-Space Models

A probabilistic state-space model is a pair of stochastic processes, the states, $X \colon \mathbb{T}_X \mapsto \mathbb{X} \subseteq \mathbb{R}^d$, and the measurements, $Y \colon \mathbb{T}_Y \mapsto \mathbb{Y} \subseteq \mathbb{R}^m$. The sets $\mathbb{T}_X$ and $\mathbb{T}_Y$ are referred to as the index sets of X and Y, respectively, and it is assumed that $\mathbb{T}_Y \subseteq \mathbb{T}_X$ In this thesis, there are two cases that are considered. The first case is continuous-discrete time state-space models for which the index sets are $\mathbb{T}_X = [0, T]$ for some $\mathbb{R} \ni T > 0$ and $\mathbb{T}_Y = \{t_n\}_{n=1}^N$, $0 \leq t_1 < t_2 < \ldots < t_N = T$. The second case is discrete time state-space models for which $\mathbb{T}_X = \{n\}_{n=0}^N$ and $\mathbb{T}_Y = \{n\}_{n=1}^N$. Throughout this chapter and the rest of the thesis, Assumption 1 is a standing assumption for any state-space model (X, Y).

Assumption 1. *For the state-space model (X, Y) the following holds:*

- *X is a Markov process.*
- *The measurements $\{Y(\tau)\}_{\tau \in \mathbb{T}_Y}$[1] are conditionally independent given the states $\{X(\tau)\}_{\tau \in \mathbb{T}_Y}$, respectively.*
- *For every $\tau \in \mathbb{T}_Y$*
$$Y(\tau) \mid X \sim p_{Y(\tau)|X(\tau)}\big(y \mid x(\tau)\big).$$

This chapter begins by discussing continuous-discrete-time probabilistic state-space models in Section 3.1. The essential concept is that of stochastic differential equations, for which, for the present purposes, the most important properties are reviewed. Some examples of continuous-discrete state-space models are also given.

In Section 3.2, discrete-time probabilistic state-space models are reviewed. This part of the chapter requires less effort to go through than the former and mostly consists of classifications of different model structures that are commonly used.

---

[1] In the discrete time case a stochastic process X evaluated at τ will be written as X_τ rather than $X(\tau)$.

3.1 Continuous-Discrete-Time State-Space Models

The class of continuous-discrete-time state-space models considered herein can be specified by

$$X(0) \sim p_{X(0)}(x), \tag{3.1a}$$

$$\mathrm{d}X(t) = a\big(t, X(t)\big)\,\mathrm{d}t + \sigma\big(t, X(t)\big)\,\mathrm{d}W(t), \tag{3.1b}$$

$$Y(t_n) \mid X \sim p_{Y(t_n)\mid X(t_n)}(t_n, y \mid x), \tag{3.1c}$$

where $a \colon [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ is a *drift function*, $\sigma \colon [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}^{d\times k}$ is a *diffusion matrix*, and $W(t)$ is a standard *Wiener process* on $\mathbb{R}^k$. Furthermore, the matrix $Q\big(t, X(t)\big) = \sigma\big(t, X(t)\big)\sigma^\mathsf{T}\big(t, X(t)\big)$ is referred to as the *instantaneous process noise covariance rate*. Equation (3.1b) is referred to as a stochastic differential equation, which requires some care to define properly. A brief review of stochastic differential equations is given in the following.

3.1.1 Stochastic Differential Equations

In this section an overview of stochastic differential equations is given. The purpose is not to give an entirely stringent account but rather introduce the concepts that are needed to apprehend the contributions of this thesis. For a more comprehensive treatment the reader is referred to the introductory textbooks such that Øksendal (2003), Särkkä and Solin (2019), or for a more technical exposition to (Karatzas and Shreve, 1988, Rogers and Williams, 2000).

The Wiener Process

The basic building block of continuous-time stochastic models is the Wiener process, which is specified in Definition 1.

Definition 1 (Wiener Process). *A stochatic process* $\{W(t)\}_{t\geq 0}$ *is said to be a* Wiener process *if the following conditions are satisfied:*

1. $W(0) = 0$ *with probability 1.*
2. *If* $0 \leq t_0 < t_1 < \cdots < t_N < \infty$ *then the increments* $W(t_n) - W(t_{n-1})$, $n = 1, \dots, N$ *are mutually independent.*
3. $W(t) - W(s) \sim \mathcal{N}\big(0, t-s\big)$ *for any* $t > s \geq 0$.

There exists a version of the Wiener process that is continuous almost surely, hence it defines a probability model on the space of continuous functions (Øksendal, 2003). Moreover, for any finite grid $\{t_n\}_{n=1}^N$, the variables $\{W(t_n)\}_{n=1}^N$ are jointly Gaussian distributed. That is, the Wiener process is a

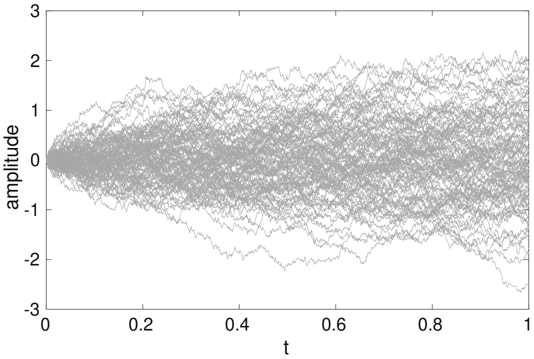

Figure 3.1. 100 independent realisations of a Wiener process.

Gaussian process (Rasmussen and Williams, 2006). A Gaussian process is characterised by its mean and covariance functions, which for the Wiener process are given by

$$\mathbb{E}[W(t)] = 0,$$

$$\mathbb{C}[W(t), W(s)] = t \wedge s.$$

Example realisations of a Wiener process are shown in Figure 3.1.

Stochastic Differential Equations

The Wiener process can be used to construct stochastic differential equations. That is, the Wiener process W can be used to define the stochastic process X by the following integral equation

$$X(t) = X(0) + \int_0^t a\big(s, X(s)\big)\,\mathrm{d}s + \int_0^t \sigma\big(s, X(s)\big)\,\mathrm{d}W(s). \tag{3.3}$$

The stochastic integral equation in Equation (3.3) is often written in differential form as a short-hand

$$\mathrm{d}X(t) = a\big(t, X(t)\big)\,\mathrm{d}t + \sigma\big(t, X(t)\big)\,\mathrm{d}W(t). \tag{3.4}$$

It is also most commonly referred to as a stochastic differential equation rather than a stochastic integral equation. It requires some care to properly define what is meant by Equation (3.3). The mischievous term is $\int_0^t \sigma\big(s, X(s)\big)\,\mathrm{d}W(s)$ for which there are two popular definitions due to Itô and Stratonovich (Øksendal, 2003). Herein the Itô interpretation is used unless otherwise stated. This breaks the ordinary chain rule and instead $X(t)$ as given by Equation (3.4) transforms according to Itô's formula, which is given in Lemma 3.

Lemma 3 (Itô's formula)**.** *Let $\phi\colon \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ be a function that is at least once differentiable in the first argument and at least twice differentiable* *in the second argument and $X(t)$ satisfies the Itô stochastic differential equation (3.4). Then $\phi(t,X(t))$ is an Itô process with Itô differential given by*

$$\mathrm{d}\phi(t,X(t)) = (\partial_t + \mathscr{A})\phi(t,X(t))\,\mathrm{d}t + \mathscr{B}\phi(t,X(t))\,\mathrm{d}W(t), \tag{3.5}$$

where the operators $\mathscr{A}$ and $\mathscr{B}_{i,j}$ are defined by

$$\mathscr{A}\phi(t,x) = \sum_i \partial_i\phi(t,x)a_i(t,x) + \frac{1}{2}\sum_{i,j}\partial^2_{ij}\phi(t,x)Q_{i,j}(t,x), \tag{3.6a}$$

$$[\mathscr{B}\phi(t,x)]_j = \sum_i \partial_i\phi(t,x)\sigma_{i,j}(t,x). \tag{3.6b}$$

The operator $\mathscr{A}$ appearing in Equation (3.6a) is the *generator* associated with the solution of Equation (3.4) and the operator $\partial_t + \mathscr{A}$ appearing in Equation (3.5) is sometimes referred to as the *generalised generator* (Särkkä and Solin, 2019, Chapter 5). The importance of the generalised generator is evident by Lemma 3, as it defines how the drift of a stochastic differential equation changes under smooth transformations. For time-homogeneous transformations, the generalised generator $\partial_t + \mathscr{A}$ reduces to the generator $\mathscr{A}$.

Remark 5. *In order for the stochastic differential equation in Equation (3.3) to have a unique solution, some conditions on a and σ are required. Herein these considerations are omitted and Equation (3.3) is always assumed to have a unique solution in some suitable sense. For a discussion on this matter the reader is referred to Karatzas and Shreve (1988).*

Another important property of the solution (in the sense of Itô) $X(t)$ of Equation (3.3) is that $X(t)$ is a Markov process, which is defined in the following.

Definition 2 (Markov process). *A continuous-time stochastic process is a Markov process if the discrete-time process $\{X(t_n)\}_{n=1}^N$ satisfies the Markov property for all $t_1 < t_2 < \ldots < t_N$. That is, if $1 \leq n' < n \leq N$*

$$p_{X(t_n)\mid\{X(t_l)\}_{l=1}^{n-1}}\big(x(t_n) \mid \{x(t_l)\}_{l=1}^{n'}\big) = p_{X(t_n)\mid X(t_{n'})}\big(x(t_n) \mid x(t_{n'})\big).$$

The Markov property is often characterised as *"the future is independent of the past given the present"* and is heavily exploited to develop efficient method for inference in the state-space model given by Equation (3.1).

The Fokker–Planck Equation

The stochastic differential equation in Equation (3.3) defines a generative model for $X(t)$ in the sense that from a Wiener process, $X(t)$ can be computed by solving Equation (3.3). However, it is fruitful to characterise the probability density imposed on $X(t)$ by $W(t)$. For this purpose, the generator $\mathscr{A}$ is crucial, or rather its formal adjoint on $\mathscr{L}_2(\mathbb{R}^d)$. That is, the probability density associated with $X(t)$ evolves according to the Fokker–Planck equation[2], which is given by Theorem 2 (see e.g., Särkkä and Solin 2019, Chapter 5).

Theorem 2 (The Fokker–Planck equation)**.** *Let $X(t)$ be the solution to Equation (3.4) and assume the probability density $p(t,x)$ for $X(t)$ exists and is supported on $\mathbb{R}^d$. Then $p(t,x)$ satisfies the following partial differential equation*

$$\dot{p}(t,x) = \mathscr{A}^\star p(t,x)$$

where $\mathscr{A}^\star$ is the adjoint of $\mathscr{A}$ as defined in Equation (3.6a) and it is given by

$$\mathscr{A}^\star p(t,x) = -\sum_i \partial_i [a_i(t,x)p(t,x)] + \frac{1}{2}\sum_{i,j} \partial^2_{i,j}\left[Q_{i,j}(t,x)p(t,x)\right].$$

Discretisation of Stochastic Differential Equations
An important tool when working with stochastic differential equations is discretisation, which entails finding a discrete-time representation of the original stochastic differential equation on some grid $0 = t_0 < t_1 < \ldots < t_N = T$. An exact probabilistic description of $\{X(t_n)\}_{n=0}^N$ is formally given by solving the Fokker–Planck equation on the intervals $[t_{n-1}, t_n]$ for $n = 1, \ldots, N$ with initial conditions $p(t_{n-1}, x) = \delta(x - x(t_{n-1}))$. For an affine model, $a(t,x) = A(t)x + b(t)$ and $\sigma(t,x) = \sigma(t)$, the solution of the Fokker–Planck equation on the interval $[t, t+h]$ with initial condition $\delta(x - x(t))$ is given by (Särkkä and Solin, 2019, Chapter 6)

$$p_{X(t+h)|X(t)}(x \mid x(t)) = \mathscr{N}\left(x; A_d(t+h \mid t)x(t) + b_d(t+h \mid t), Q_d(t+h \mid t)\right),$$

where

$$\partial_h A_d(t+h \mid t) = A(t)A_d(t+h|t) \quad A_d(t \mid t) = \mathrm{I}, \tag{3.7a}$$

$$b_d(t+h \mid t) = \int_t^{t+h} A_d(t+h|\tau)b(\tau)\,\mathrm{d}\tau, \tag{3.7b}$$

$$Q_d(t+h \mid t) = \int_t^{t+h} A_d(t+h|\tau)Q(\tau)A_d^\top(t+h|\tau)\,\mathrm{d}\tau. \tag{3.7c}$$

The Fokker–Planck equation is intractable in general wherefore approximate discretisations are important, examples of which are Itô–Taylor series expansions and stochastic Runge–Kutta methods (Särkkä and Solin, 2019, Chapter 8). The former class of methods are derived by iteratively invoking Lemma 3 on $a(t, X(t))$ and $\sigma(t, X(t))$. After a single application, the following representation is retrieved

$$X(t+h) = X(t) + a\left(t, X(t)\right)h + \sigma\left(t, X(t)\right)\Delta_h W(t) + R_X(t, t+h),$$

---

[2] The Fokker-Planck equation is also known as the Kolmogorov forward equation.

where $\Delta_h W(t) = W(t+h) - W(t)$ is a Wiener increment of size h and $R_X(t, t+h)$ is a remainder term. Neglecting the remainder term then yields the Euler–Maruyama scheme

$$X(t+h) = X(t) + a\big(t, X(t)\big)h + \sigma\big(t, X(t)\big)\Delta_h W(t),$$

which for present purposes is the only needed approximate discretisation scheme. For a more thorough exposition on discretisation methods the reader is referred to Kloeden and Platen (2013).

Basis Expansions of the Wiener Process

The Wiener process on $\mathbb{R}^k$ restricted to the interval $[0, T]$ can be expressed in terms of a basis of $\mathscr{L}_2([0, T])$. More, specifically if $\{\phi_l\}_{l=1}^{\infty}$ is an orthonormal basis of $\mathscr{L}_2([0, T])$, then the Wiener process $\{W(t)\}_{0 \leq t \leq T}$ has the following Fourier expansion (Luo, 2006)

$$W(t) = \sum_{l=1}^{\infty} U_l \int_0^t \phi_l(\tau)\, \mathrm{d}\tau, \tag{3.8}$$

and the coefficients are given by

$$U_l = \int_0^T \phi_l(\tau)\, \mathrm{d}W(\tau),$$

where $\mathbb{E}[U_l] = 0$, $\mathbb{C}[U_i, U_j] = \delta_{ij}\mathrm{I}$. The series expansion in Equation (3.8) can be truncated at the Lth term to give a finite dimensional approximation of the Wiener process

$$\hat{W}_L(t) = \sum_{l=1}^{L} U_l \int_0^t \phi_l(\tau)\, \mathrm{d}\tau. \tag{3.9}$$

Furthermore, $\hat{W}_L$ converges in the mean square to W at a rate of L if the trigonometric basis is selected (Luo, 2006, Theorem 2.1)

$$\phi_l(t) = \begin{cases} \frac{1}{\sqrt{T}}, & l = 1, \\ \sqrt{\frac{2}{T}} \cos\left(\frac{(l-1)\pi t}{T}\right), & l > 1. \end{cases}$$

An example of the approximation for one realisation of the Wiener process is shown in Figure 3.2.

Furthermore, Equation (3.9) gives an approximation to the Wiener increment $\mathrm{d}W(t)$ on the interval $[0, T]$ by

$$\mathrm{d}W(t) \approx \mathrm{d}\hat{W}_L(t) = \sum_{l=1}^{L} U_l \phi_l(t)\, \mathrm{d}t,$$

which in turn gives an approximation to the stochastic differential equation

$$\mathrm{d}X(t) = a\big(t, X(t)\big)\, \mathrm{d}t + \sigma\big(t, X(t)\big)\, \mathrm{d}W(t) \tag{3.10}$$

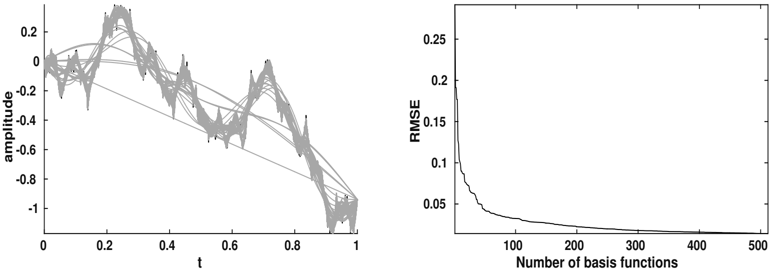

Figure 3.2. The convergence of $\hat{W}_L$ to W for $L = 1,\ldots,2^8$. $\hat{W}_L$ is shown to the left and the root mean square error (RMSE) is shown to the right.

on the interval $[0,T]$ by (Särkkä and Solin, 2019, Section 9.8)

$$\mathrm{d}\hat{X}_L(t) = a\big(t,\hat{X}_L(t)\big)\,\mathrm{d}t + \sigma\big(t,\hat{X}_L(t)\big)\sum_{l=1}^{L} U_l\phi_l(t)\,\mathrm{d}t. \tag{3.11}$$

While $\hat{W}_L$ converges to W as $L \to \infty$, examining in what sense the solution of Equation (3.11) $\hat{X}_L$ converges to the solution of Equation (3.10) X as $L \to \infty$ requires some care. This has been examined in the one dimensional case, where it turns out that $\hat{X}_L$ converges to X if Equation (3.10) is interpreted in Stratonovich sense (Wong and Zakai, 1965). For convergence in higher dimensions the reader is referred to the appendix of Lyons et al. 2014 and references therein. The approximation in Equation (3.11) is used in Publication VI to develop a series approximation for Gaussian smoothers.

3.1.2 Examples of Continuous-Discrete-Time State-Space Models

In this section, some of the more prominent classes of continuous-discrete state-space models are reviewed together with some examples.

Affine Gaussian Models

Affine Gaussian state-space models are of the following form:

$$\mathrm{d}X(t) = A(t)X(t)\,\mathrm{d}t + b(t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W(t), \tag{3.12a}$$

$$Y(t_n) \mid X \sim \mathcal{N}\big(C(t_n)x(t_n) + d(t_n), R(t_n)\big). \tag{3.12b}$$

If A, b, σ, C, d, and R are constant the model is said to be *affine time-invariant* (ATI), if additionally $b = 0$ and $d = 0$ then the model is said to be *linear time-invariant* (LTI). On the contrary, if A, b, σ, C, d, and R are time-varying then the model is said to be *affine time-variant* (ATV), or *linear time-variant* (LTV) ($b = 0$ and $d = 0$). An important example of a LTI system is the q times integrated Wiener process on $\mathbb{R}^d$ with linear

Gaussian measurements,

$$\mathrm{d}X^{(i)}(t) = \begin{cases} X^{(i+1)}(t)\,\mathrm{d}t, & i = 1,\dots,q, \\ \Gamma^{1/2}\,\mathrm{d}W(t), & i = q+1, \end{cases} \tag{3.13a}$$

$$Y(t_n) \mid X \sim \mathcal{N}\big(Cx(t_n), R\big), \tag{3.13b}$$

where the sub-vectors $X^{(i)}(t) \in \mathbb{R}^d$ of the complete state $X(t)$ are $q+1-i$ times iterated integrals of the Wiener process $\Gamma^{1/2}W(t) \in \mathbb{R}^d$, and $\Gamma^{1/2} \in \mathbb{R}^{d\times d}$ is the symmetric square root of some positive definite matrix Γ. The model matrices associated with Equation (3.13a) are given by

$$A = \left(\sum_{i=1}^{q} \mathrm{e}_i \mathrm{e}_{i+1}^{\mathsf{T}}\right) \otimes \mathrm{I}_d,$$

$$\sigma = \mathrm{e}_{q+1} \otimes \Gamma^{1/2},$$

where e_i are canonical basis vectors in $\mathbb{R}^d$ and $\otimes$ denotes Kronecker's product. If R is formally set to zero and the measurement matrix is given by

$$C = -\mathrm{e}_1^{\mathsf{T}} \otimes \Lambda + \mathrm{e}_2^{\mathsf{T}} \otimes \mathrm{I}_d,$$

then Equation (3.13) is a model for the solution of the following differential equation (Tronarp et al., 2019):

$$\dot{y}(t) = \Lambda y(t).$$

This type of model was used to develop probabilistic solvers for ordinary differential equations in Publication III, though in general non-linear measurement models are required. An example realisation of the q times integrated Wiener process is shown in Figure 3.3.

Remark 6. *In the tracking literature it is common to take $q = 1$, in which case the dynamic model in Equation* (3.13a) *is known as the* nearly constant velocity model (CV) *or* Wiener velocity model *(Li and Jilkov, 2003).*

Conditionally Gaussian Models

Conditionally Gaussian models are of the following form.

$$\mathrm{d}X(t) = a\big(t, X(t)\big)\,\mathrm{d}t + \sigma\big(t, X(t)\big)\,\mathrm{d}W(t), \tag{3.15a}$$

$$Y(t_n) \mid X \sim \mathcal{N}\big(c(t_n, x(t_n)), R(t_n, x(t_n))\big). \tag{3.15b}$$

Conditionally Gaussian models might be a poor choice of words because $X(s) \mid X(t), \quad s > t$ is not Gaussian distributed in general, though infinitesimally it is correct in the sense that the following holds formally.

$$X(t + \mathrm{d}t) = X(t) + a\big(t, X(t)\big)\,\mathrm{d}t + \sigma\big(t, X(t)\big)\,\mathrm{d}W(t).$$

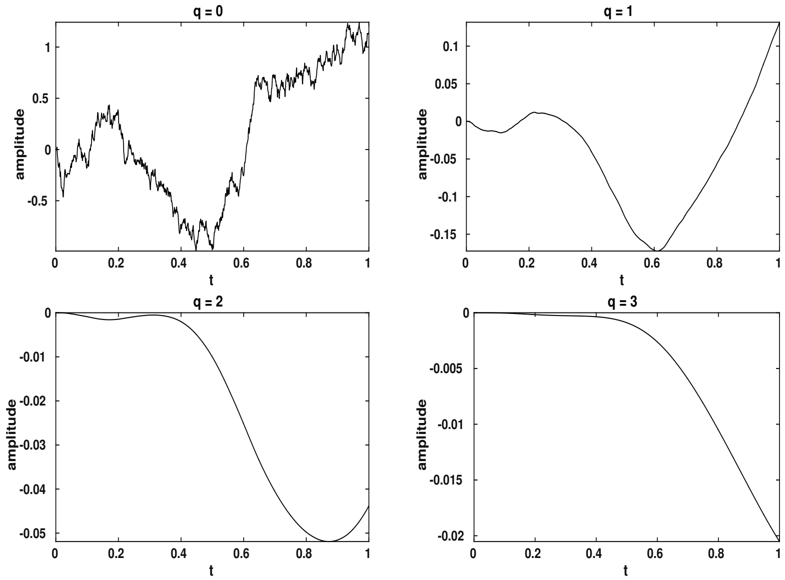

Figure 3.3. One realisation of the q times integrated Wiener process ($X^{(1)}$) on $\mathbb{R}$ for $q = 0, 1, 2, 3$.

An example of a continuous-discrete conditionally Gaussian model is the radar tracked coordinates turn model. The dynamic model is given by

$$\begin{aligned}
\mathrm{d}P_X(t) &= \dot{P}_X(t)\,\mathrm{d}t,\\
\mathrm{d}P_Y(t) &= \dot{P}_Y(t)\,\mathrm{d}t,\\
\mathrm{d}P_Z(t) &= \dot{P}_Z(t)\,\mathrm{d}t,\\
\mathrm{d}\dot{P}_X(t) &= -\Psi(t)\dot{P}_Y(t)\,\mathrm{d}t + \sigma_X\,\mathrm{d}W_X(t),\\
\mathrm{d}\dot{P}_Y(t) &= \Psi(t)\dot{P}_X(t)\,\mathrm{d}t + \sigma_Y\,\mathrm{d}W_Y(t),\\
\mathrm{d}\dot{P}_Z(t) &= \sigma_Z\,\mathrm{d}W_Z(t)\\
\mathrm{d}\Psi(t) &= \sigma_\Psi\,\mathrm{d}W_\Psi(t),
\end{aligned}$$

where (P_X, P_Y, P_Z) is the position (of, e.g., an aircraft), $(\dot{P}_X, \dot{P}_Y, \dot{P}_Z)$ is the velocity vector, ψ is the turn rate, and W_X, W_Y, W_Z, and W_Ψ are mutually independent standard Wiener processes. Essentially, the coordinated turn model is a Wiener velocity model in the Z-direction with the rate of change in velocity determined by $\sigma_Z \in \mathbb{R}_+$. In the X-Y plane, this is a circular motion around origin with angular rate Ψ, which is perturbed by the Wiener processes W_X and W_Y with the magnitude of the perturbations determined by $\sigma_X, \sigma_Y \in \mathbb{R}_+$. Lastly, the turn rate is subject to a perturbation by W_Ψ with its magnitude determined by $\sigma_\Psi \in \mathbb{R}_+$. By denoting the complete state vector by X, this can be written more compactly as

$$\mathrm{d}X(t) = a\big(X(t)\big)\,\mathrm{d}t + \sigma\,\mathrm{d}W(t),$$

where

$$\begin{aligned}
X^\top &= \begin{pmatrix} P_X & P_Y & P_Z & \dot{P}_X & \dot{P}_Y & \dot{P}_Z & \Psi \end{pmatrix},\\
W^\top &= \begin{pmatrix} W_X & W_Y & W_Z & W_\Psi \end{pmatrix},\\
a(X) &= \begin{pmatrix} \dot{P}_X & \dot{P}_Y & \dot{P}_Z & -\Psi\dot{P}_Y & \Psi\dot{P}_X & 0 & 0 \end{pmatrix},\\
\sigma^\top &= \begin{pmatrix} 0 & 0 & 0 & \sigma_X & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & \sigma_Y & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & \sigma_Z & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_\psi \end{pmatrix}.
\end{aligned}$$

As the term ”radar-tracked“ might suggest, the state is measured by a radar. That is, measurements are taken of the range $\sqrt{P_X^2 + P_Y^2 + P_Z^2}$, the azimuth angle $\tan^{-1}\left(P_Y/P_X\right)$, and the elevation angle $\tan^{-1}\left(P_Z/\sqrt{P_X^2 + P_Y^2}\right)$ subject to independent Gaussian measurement errors for each measured quantity. This can be written as

$$\begin{aligned}
&c^\top(X(t_n)) = \left(\sqrt{P_X(t_n)^2 + P_Y(t_n)^2 + P_Z(t_n)^2} \quad \tan^{-1}\frac{P_Y(t_n)}{P_X(t_n)} \quad \tan^{-1}\frac{P_Z(t_n)}{\sqrt{P_X^2(t_n) + P_Y^2(t_n)}} \right),\\
&\quad V(t_n) \sim \mathcal{N}\big(0, \mathrm{diag}(\sigma_R^2, \sigma_\Phi^2, \sigma_\Theta^2)\big),\\
&\quad Y(t_n) = c\big(X(t_n)\big) + V(t_n).
\end{aligned}$$

Consequently, the complete state-space model is given by

$$\begin{aligned}
\mathrm{d}X(t) &= a\big(X(t)\big)\,\mathrm{d}t + \sigma\,\mathrm{d}W(t),\\
Y(t_n) \mid X &\sim \mathcal{N}\Big(c(X(t_n)), \mathrm{diag}(\sigma_R^2, \sigma_\Phi^2, \sigma_\Theta^2)\Big).
\end{aligned}$$

This is a slightly simpler version of a dynamic model that was used in Publication I. An example realisation of the radar tracked coordinated turn model is shown in Figure 3.4.

A Spherical State-Space

In, for example, reference vector tracking the state-space is identified with the unit sphere in $\mathbb{R}^3$, $\mathbb{X} = \mathbb{S}^2$. A simple model for a stochastic process on $\mathbb{S}^2$ is given by

$$\mathrm{d}X(t) = -\breve{\Omega}(t) \times X(t)\,\mathrm{d}t - \gamma^2 X(t)\,\mathrm{d}t + \gamma X(t) \times \mathrm{d}W(t), \tag{3.20}$$

where $\breve{\Omega}(t)$ is a deterministic angular velocity vector and $\gamma > 0$ is a scalar diffusion constant. This model was used in Publication V to model the

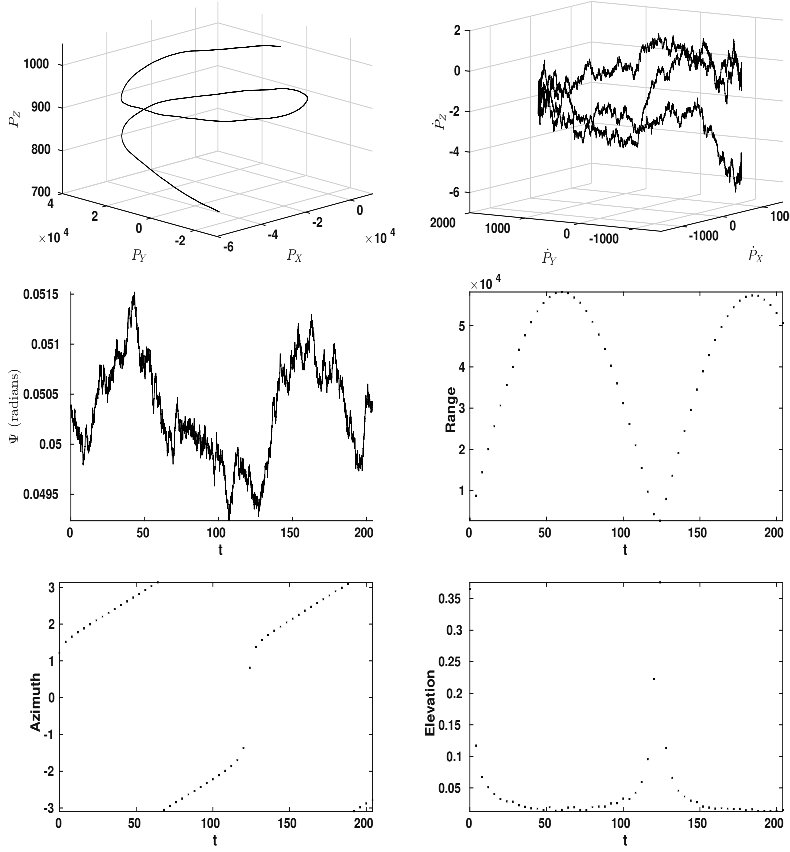

Figure 3.4. Example realisation of the radar tracked coordinated turn model.

evolution of the local gravity vector with measured angular velocity $\breve{\Omega}(t)$ subject to noise. If $\breve{\Omega} = 0$ and $\gamma = 1$ then Equation (3.20) reduces to the Wiener process on $\mathbb{S}^2$ (Price and Williams, 1983, Van Den Berg and Lewis, 1985). Thus if X models the local gravity vector of some sensor platform then a suitable state-space model is given by

$$\mathrm{d}X(t) = -\breve{\Omega}(t) \times X(t)\,\mathrm{d}t - \gamma^2 X(t)\,\mathrm{d}t + \gamma X(t) \times \mathrm{d}W(t),$$

$$Y(t_n) \mid X \sim \mathcal{N}\big(gx(t_n), \sigma^2 \mathrm{I}\big),$$

where $Y(t_n)$ are accelerometer readings and g is the local gravity constant. An example realisation of Equation (3.20) is shown in Figure 3.5.

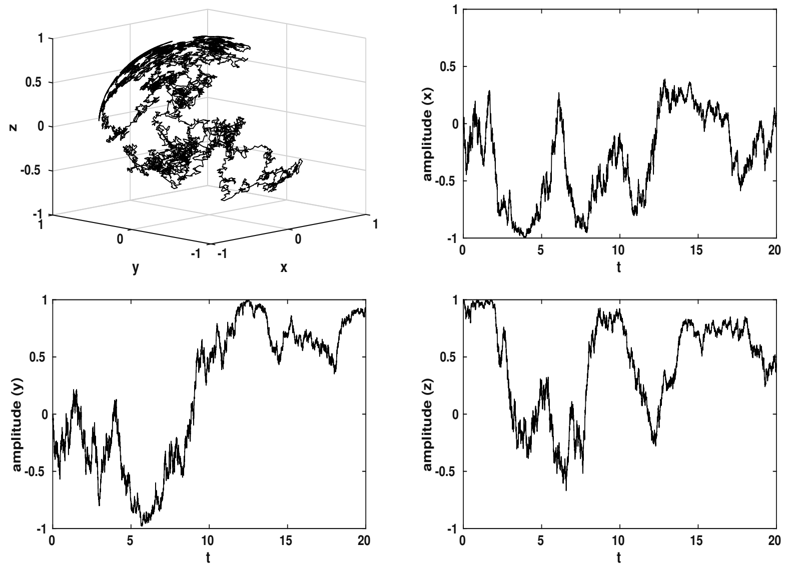

Figure 3.5. One realisation of the stochastic rotation model in Equation (3.20) with $\breve{\Omega} = 0$ (Spherical Wiener process).

3.2 Discrete-Time State-Space Models

Discrete-time Markov models are defined via a *Markov kernel* or *transition density*. Assuming the stochastic process $\{X_n\}_{n \in \mathbb{N}_0}$ is a Markov process, then similarly to the continuous-time case (see Definition 2) the following holds for $n' \leq n$.

$$p_{X_n \mid \{X_l\}_{l=0}^{n'}}\left(x_n \mid \{x_l\}_{l=1}^{n'}\right) = p_{X_n \mid X_{n'}}(x_n \mid x_{n'}).$$

In particular, if $n' = n - 1$, then

$$p_{X_n \mid \{X_l\}_{l=0}^{n-1}}\left(x_n \mid \{x_l\}_{l=0}^{n-1}\right) = p_{X_n \mid X_{n-1}}(x_n \mid x_{n-1})$$

and $p_{X_n \mid X_{n-1}}(x_n \mid x_{n-1})$ is said to be the transition density (at time n). The transition densities $p_{X_n \mid X_{n-1}}$, $n = 1, \ldots, N$ fully characterise the process X. Consequently, due to Assumption 1, probabilistic state-space models in discrete time are fully specified by the following probability densities:

$$\begin{aligned} X_0 &\sim p_{X_0}(x), \\ X_n \mid X_{n-1} &\sim p_{X_n \mid X_{n-1}}(x \mid x_{n-1}), \\ Y_n \mid X_n &\sim p_{Y_n \mid X_n}(y \mid x_n). \end{aligned}$$

3.2.1 Examples of Discrete-Time State-Space Models

In this section, some of the more prominent classes of discrete time state-space models are reviewed together with some examples.

Affine Gaussian Models

In discrete time, affine Gaussian state-space models are of the following form.

$$\begin{aligned} X_0 &\sim \mathcal{N}(\mu_0, \Sigma_0), \\ X_n \mid X_{n-1} &\sim \mathcal{N}\big(A_n x_{n-1} + b_n, Q_n\big), \\ Y_n \mid X_n &\sim \mathcal{N}(C_n x_n + d_n, R_n). \end{aligned}$$

The distinctions between *affine time-invariant* (ATI), *affine time-variant* (ATV), *linear time-invariant* (LTI), and *linear time-variant* (LTV) are made completely analogously to the continuous time setting. Recall the q times integrated Wiener process as defined in Equation (3.13). Given a grid $\{t_n\}_{n=0}^N$ with $t = 0$ and $t_N = T$ it can be discretised exactly using the methods described in Section 3.1.1. If a uniform grid is selected, $t_n = t_{n-1} + h,\ n = 1, \ldots, N$, then the discretised model is linear time-invariant, $X(nh) = X_n$,

$$X_n \mid X_{n-1} \sim \mathcal{N}(A X_{n-1}, Q),$$

where the $d \times d$ blocks of A and Q are given by

$$\begin{aligned} A_{ij} &= \begin{cases} \mathrm{I}_d \frac{h^{j-i}}{(j-i)!}, \ i \leq j, \\ 0, \ i > j. \end{cases}, \\ Q_{ij} &= \mathrm{I}_d \frac{h^{2q+3-i-j}}{(2q+3-i-j)(q+1-i)!(q+1-j)!}. \end{aligned}$$

If the "position" of X, $X^{(1)}$, is measured with an error of covariance R, then the measurement model is given by

$$\begin{aligned} C &= \begin{pmatrix} \mathrm{I}_d & 0 & \ldots & 0 \end{pmatrix}, \\ Y_n \mid X_n &\sim \mathcal{N}\big(C X_n, R\big). \end{aligned}$$

Conditionally Gaussian Models

In discrete time, conditionally Gaussian models have the property that $X_n \mid X_{n-1}$ and $Y_n \mid X_n$ are indeed Gaussian distributed. That is, the model is given by

$$X_0 \sim \mathcal{N}(\mu_0, \Sigma_0), \tag{3.26a}$$

$$X_n \mid X_{n-1} \sim \mathcal{N}\big(a_n(x_{n-1}), Q_n(x_{n-1})\big), \tag{3.26b}$$

$$Y_n \mid X_n \sim \mathcal{N}\big(c_n(x_n), R_n(x_n)\big). \tag{3.26c}$$

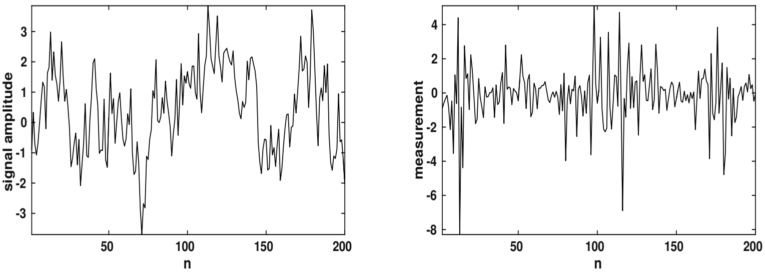

Figure 3.6. One realisation of the stochastic volatility model. The signal is on the left and the measurements are on the right.

An example of a conditionally Gaussian model is the stochastic volatility model, which is given by

$$\begin{aligned} X_n \mid X_{n-1} &\sim \mathcal{N}\big(a(X_{n-1}-m)+m, Q\big), \\ Y_n \mid X_n &\sim \mathcal{N}\big(0, \exp(X_n)\big). \end{aligned}$$

The stochastic volatility model is commonly used to model *volatility* in financial time series. That is, the variance of returns (Lindström et al., 2015). An example realisation of the stochastic volatility model is shown in Figure 3.6.

4. Inference in Probabilistic State-Space Models

The previous chapter covered the mathematical description of the latent signal X and its relation with the measurement signal Y. That is, the mathematical description of the system (X,Y) with which this thesis is concerned with. The present chapter is concerned with inferring the latent signal X given the measurements Y. For this purpose, the following two sets are defined.

$$\mathscr{Y}(\tau) = \big\{Y(s) \colon s \leq \tau\big\}, \quad \tau, s \in \mathbb{T}_X, \tag{4.1a}$$

$$\mathscr{Y}^-(\tau) = \big\{Y(s) \colon s < \tau\big\}, \quad \tau, s \in \mathbb{T}_X. \tag{4.1b}$$

The sets $\mathscr{Y}(\tau)$ and $\mathscr{Y}^-(\tau)$ is the information provided by the measurements Y about the latent signal X up to time τ and up to *just before* time τ, respectively. The complete information provided by the measurements is simply the union of all $\mathscr{Y}(\tau)$ and it is denoted by $\mathscr{Y}$. Furthermore, the sets in Equation (4.1) define two classes of conditional densities through Bayes' rule, namely

$$p\big(s,x \mid \mathscr{Y}(\tau)\big), \tag{4.2a}$$

$$p\big(s,x \mid \mathscr{Y}^-(\tau)\big). \tag{4.2b}$$

There are some important classifications of the densities in Equation (4.2) for particular choices of τ and s as listed below.

- If $s = \tau \in \mathbb{T}_Y$ then $p\big(s,x \mid \mathscr{Y}(\tau)\big)$ is a filtering density and $p\big(s,x \mid \mathscr{Y}^-(\tau)\big)$ is a (one-step ahead) prediction density.

- If $s < \tau \in \mathbb{T}_Y$ then both $p\big(s,x \mid \mathscr{Y}(\tau)\big)$ and $p\big(s,x \mid \mathscr{Y}^-(\tau)\big)$ are smoothing densities. Unless τ is the smallest element in $\mathbb{T}_Y$, in which case $p\big(s,x \mid \mathscr{Y}^-(\tau)\big)$ is a prediction density.

- If $s > \tau \in \mathbb{T}_Y$ then both $p\big(s,x \mid \mathscr{Y}(\tau)\big)$ and $p\big(s,x \mid \mathscr{Y}^-(\tau)\big)$ are prediction densities.

Herein the concern is with inference on a fixed interval. The index sets are given by $\mathbb{T}_X = [0,T]$ and $\mathbb{T}_Y = \{t_n\}_{n=1}^N$, $0 \leq t_1 < \ldots t_N = T$ or $\mathbb{T}_X = \{n\}_{n=0}^N$ and $\mathbb{T}_Y = \{n\}_{n=1}^N$ for the continuous-discrete time and the discrete time inference problems, respectively.

4.1 Bayesian Inference in Continuous-Discrete Models

In this section inference in continuous-discrete time models is discussed. It starts off with describing the formal solution to the filtering and smoothing problem. The solution to the former problem can be expressed using the the Fokker–Planck equation and Bayes' rule, while the latter problem also requires some results from Leondes et al. (1970) or equivalently Anderson (1972). The discussion of the formal solution is then followed by inference in affine Gaussian models for which inference is done via the Kalman filter (Kalman, 1960) and the Rauch–Tung–Striebel smoother (Rauch et al., 1965). This section concludes by discussing approximate strategies for conditionally Gaussian models with emphasis on the assumed density approach (Maybeck, 1979,1982,1982) and its relation to the projection methods (Brigo et al., 1999, Koyama, 2018).

4.1.1 The Formal Solution

The formal solution of the continuous-discrete time inference problem involves finding the filtering and smoothing densities associated with the following state-space model.

$$X(0) \sim p_{X(0)}(x), \tag{4.3a}$$

$$\mathrm{d}X(t) = a\big(t, X(t)\big)\,\mathrm{d}t + \sigma\big(t, X(t)\big)\,\mathrm{d}W(t), \tag{4.3b}$$

$$Y(t_n) \mid X \sim p_{Y(t_n)\mid X(t_n)}\big(t_n, y \mid x\big), \tag{4.3c}$$

What is meant by formal here is that precise mathematical relations between the filtering densities, likelihoods, and smoothing densities are given. These relations need not be appropriate for implementation on a computer system but do define the bullseye for approximate methods.

The Formal Filtering Relations

The filtering density is denoted by $p_F(x,t) = p\big(x, t \mid \mathscr{Y}(t)\big)$. In the intervals between measurements it is governed by the Fokker–Planck equation

$$\dot{p}_F(t,x) = \mathscr{A}^\star p_F(t,x), \quad t \in [t_{n-1}, t_n). \tag{4.4}$$

The filtering density is a so-called cádlág function (left continuous with right limits). That is, the half-open integration interval in Equation (4.4) indicate that, strictly speaking, $p_F(t_n,x)$ is not the object that is computed, but rather the prediction density $p_F(t_n^-,x) = \lim_{s\uparrow t_n} p_F(t_n,x)$. The filtering density at time t_n is then formally computed via Bayes' rule, using the prediction density and the likelihood.

$$p_F(t_n,x) = \frac{L(t_n,x)p_F(t_n^-,x)}{\int_{\mathbb{X}} L(t_n,x)p_F(t_n^-,x)\,\mathrm{d}x}. \tag{4.5}$$

Recall that the likelihood is given by $L(t_n,x) = p_{Y(t_n)|X(t_n)}(t_n, y(t_n) \mid x)$.

The Formal Smoothing Relations

There's a formal relationship between the smoothing density $p_S(t,x) = p(x,t \mid \mathscr{Y}(t_N))$ and the filtering density $p_F(t,x)$, which similarly to the prediction in Equation (4.4) takes the form of a partial differential equation. Before arriving at this conclusion, it is fruitful to give Proposition 1 (Leondes et al., 1970, Eqs. (25) and (26)).

Proposition 1. *Let the stochastic processes X and Y be governed by Equation (4.3). Assume that the filtering and smoothing probability densities exist and are supported on the entirety of $\mathbb{R}^d$. Then the mean of a smooth function $\phi(X(t))$ satisfies the following relation under the smoothing distribution*

$$\partial_t \mathbb{E}\left[\phi(X(t)) \mid \mathscr{Y}(t_N)\right] = \mathbb{E}\left[\mathscr{K}_1\phi(X(t)) \mid \mathscr{Y}(t_N)\right],$$

where the operator $\mathscr{K}_1$ is defined by

$$\begin{aligned}\mathscr{K}_1\phi(x) = &\sum_i a_i(t,x)\partial_i\phi(x) - \frac{1}{2}\sum_{i,j} Q_{ij}(t,x)\partial_{ij}^2\phi(x) \\ &- \sum_{i,j} \partial_i\phi(x)\partial_j Q_{ij}(t,x) \\ &- \sum_{i,j} Q_{ij}(t,x)\partial_i\phi(x)\frac{\partial_j p_F(t,x)}{p_F(t,x)}.\end{aligned} \tag{4.6}$$

The evolution of the smoothing density given in Theorem 3 follows directly from Proposition 1 and the method of adjoints on $\mathscr{L}_2(\mathbb{R}^d)$, which gives a partial differential equation for the smoothing density p_S.[1]

Theorem 3. *Let the stochastic processes X and Y be governed by Equation (4.3). Assume that the filtering and smoothing probability densities exist and are supported on the entirety of $\mathbb{R}^d$. Then the smoothing density is given by*

$$\begin{aligned}\dot{p}_S(t,x) = &-\sum_i \partial_i\left[a_i(t,x)p_S(t,x)\right] - \frac{1}{2}\sum_{i,j}\partial_{ij}^2\left[Q_{i,j}(t,x)p_S(t,x)\right] \\ &+ \sum_{i,j}\partial_i\left[\partial_j Q_{i,j}(t,x)p_S(t,x)\right] + \sum_{i,j}\partial_i\left[Q_{i,j}(t,x)p_S(t,x)\partial_j \log p_F(t,x)\right].\end{aligned} \tag{4.7}$$

---

[1] A partial differential equation for the smoothing density is also given in Equation (28) of Leondes et al. (1970). Unfortunately, in the original reference there is a sign error in the first term on the right-hand side.

This can be written more compactly as

$$\dot{p}_S(t,x) = \mathcal{K}_1^\star p_S(t,x),$$

where the adjoint $\mathcal{K}_1^\star$ *of* $\mathcal{K}_1$ *is given by*

$$\begin{aligned}\mathcal{K}_1^\star \phi(x) = &-\sum_i \partial_i \left[a_i(t,x)\phi(x)\right] - \frac{1}{2}\sum_{i,j} \partial_{ij}^2 \left[Q_{i,j}(t,x)\phi(x)\right] \\ &+ \sum_{i,j} \partial_i \left[\partial_j Q_{i,j}(t,x)\phi(x)\right] + \sum_{i,j} \partial_i \left[Q_{i,j}(t,x)\phi(x)\partial_j \log p_F(t,x)\right].\end{aligned} \tag{4.8}$$

Equation (4.7) *is solved backwards from* $t = t_N$ *with terminal condition* $p_S(t_N,x) = p_F(t_N,x)$.

Remark 7. Anderson (1972) *presents another expression for the evolution of the smoothing density in terms of the generator* $\mathscr{A}$ *and its adjoint* $\mathscr{A}^\star$, *namely*

$$\dot{p}_S(t,x) = \mathscr{A}^\star[p_F](t,x) - p_F(t,x)\mathscr{A}\left[\frac{p_S}{p_F}\right](t,x).$$

4.1.2 Inference in Affine Models

It is rarely the case that the filtering and smoothing densities are in some finite dimensional space, which makes their computation intractable in general. However, there are some exceptions such as, Benĕs and Benĕs–Daum problems (Beneš, 1981, Daum, 1984). Though, the most important of such cases arise from the affine Gaussian model, which reduces the inference problem to manipulations of the joint two first moments of X and Y. Then the filtering densities can be computed with the *Kalman filter* (Kalman, 1960) and the smoother densities with the *Rauch–Tung–Striebel smoother* (Rauch et al. 1965, see also Striebel 1965).

Continuous-Discrete Time Kalman Filter

The continuous-discrete Kalman filter computes the filtering densities for the affine Gaussian model in Equation (3.12). If the filtering density at time t_{n-1} is given by

$$p_F(t_{n-1},x) = \mathcal{N}\left(x; \mu^F(t_{n-1}), \Sigma^F(t_{n-1})\right),$$

then the prediction density at any time $t > t_{n-1}$ will be Gaussian as well. In particular, the prediction density at time t_n, which is denoted by

$$p_F(t_n^-,x) = \mathcal{N}\left(x; \mu^F(t_n^-), \Sigma^F(t_n^-)\right),$$

is retrieved by solving the following set of ordinary differential equations

$$\begin{aligned}&\dot{\mu}^F(t) = A(t)\mu^F(t) + b(t), \\ &\dot{\Sigma}^F(t) = A(t)\Sigma^F(t) + \Sigma^F(t)A^\top(t) + Q(t), \ t \in [t_{n-1}, t_n).\end{aligned}$$

Furthermore, the filtering density at time t_n is

$$p_F(t_n,x) = \mathcal{N}\big(x;\mu^F(t_n),\Sigma^F(t_n)\big),$$

with parameters given by

$$\begin{aligned}
S(t_n) &= C(t_n)\Sigma^F(t_n^-)C^\top(t_n) + R(t_n),\\
K(t_n) &= \Sigma^F(t_n^-)C^\top(t_n)S^{-1}(t_n),\\
\mu^F(t_n) &= \mu^F(t_n^-) + K(t_n)\big(y(t_n) - C(t_n)\mu^F(t_n^-) - d(t_n)\big),\\
\Sigma^F(t_n) &= \Sigma^F(t_n^-) - K(t_n)S(t_n)K^\top(t_n),
\end{aligned}$$

which is referred to as the *Kalman update*.

Continuous-Time Rauch–Tung–Striebel Smoother
The continuous-time Rauch–Tung–Striebel smoother takes the filtering density as input to produce the set of smoothing densities on $[0,T]$. More specifically, the smoothing density is given by

$$p_S(t,x) = \mathcal{N}(x;\mu^S(t),\Sigma^S(t)),$$

and

$$\dot\mu^S(t) = A(t)\mu^S(t) + b(t) + Q(t)\big[\Sigma^F(t)\big]^{-1}\big(\mu^S(t) - \mu^F(t)\big), \tag{4.11a}$$

$$\dot\Sigma^S(t) = \Big[A(t) + Q(t)\big[\Sigma^F(t)\big]^{-1}\Big]\Sigma^S(t) + \Sigma^S(t)\Big[A(t) + Q(t)\big[\Sigma^F(t)\big]^{-1}\Big]^\top - Q(t). \tag{4.11b}$$

The ordinary differential equations in Equation (4.11) are solved backwards in time on the interval $[0,T]$ with terminal conditions $\mu^S(T) = \mu^F(T)$ and $\Sigma^S(T) = \Sigma^F(T)$. Equation (4.11) can be written more succinctly by defining the parameters

$$\begin{aligned}
A^S(t) &= A(t) + Q(t)\big[\Sigma^F(t)\big]^{-1},\\
b^S(t) &= b(t) - Q(t)\big[\Sigma^F(t)\big]^{-1}\mu^F(t),
\end{aligned}$$

which gives

$$\begin{aligned}
\dot\mu^S(t) &= A^S(t)\mu^S(t) + b^S(t),\\
\dot\Sigma^S(t) &= A^S(t)\Sigma^S(t) + \Sigma^S(t)\big[A^S(t)\big]^\top - Q(t).
\end{aligned}$$

4.1.3 Approximate Inference in Non-Affine Models

Many approaches to inference in continuous-time non-affine models proceed in a similar manner to the approximate inference methods with Gaussian priors as discussed in Section 2.1.1, which are particular instances of the assumed density approach (Maybeck, 1979,1982,1982). However, other approaches exists as well, such as sequential Monte Carlo methods (Särkkä and Sottinen, 2008), fixed-form variational Bayes (Ala-Luhtala et al., 2015, Archambeau et al., 2008, Sutter et al., 2016), expectation propagation (Cseke et al., 2016), and the projection methods (Brigo et al., 1998, 1999, Koyama, 2018). In the following some of the most prominent approaches to assumed Gaussian density estimation in non-affine models (see Eq. (3.15)) are discussed, followed by a short description of the projection methods.

The Assumed Density Approach

As the name suggests, assumed density filters and smoothers approximate the formal filtering and smoother solutions Equations (4.4), (4.5), and (4.7), under the assumption that the filtering and smoothing densities are in a particular class of densities (Maybeck, 1979,1982,1982). This class is most commonly taken to be Gaussian, though deviations exists (see, e.g., Lee 2018a,b, Tronarp et al. 2018b). From Equation (3.4) and Lemma 3, differential equations for the predictive mean and covariance can be derived[2]. They are given by

$$\partial_t \mathbb{E}\big[X(t) \mid \mathscr{Y}(t)\big] = \mathbb{E}\big[a(t,X(t)) \mid \mathscr{Y}(t)\big], \tag{4.14a}$$

$$\begin{aligned}\partial_t \mathbb{V}[X(t) \mid \mathscr{Y}(t)] &= \mathbb{E}\big[a(t,X(t))(X(t)-\mu^F(t))^\mathsf{T} \mid \mathscr{Y}(t)\big] \\ &\quad + \mathbb{E}\big[(X(t)-\mu^F(t))a(t,X(t))^\mathsf{T} \mid \mathscr{Y}(t)\big] + \mathbb{E}\big[Q(t,X(t)) \mid \mathscr{Y}(t)\big],\end{aligned} \tag{4.14b}$$

where $\mathbb{E}[\cdot \mid \mathscr{Y}(t)]$ is the expectation with respect to the exact filtering density.

The assumed Gaussian density approach operates by replacing these expectations in Equation (4.14) with expectations with respect to the Gaussian approximation of the filtering density. That is, the assumed Gaussian prediction equations are given by Approximation 6.

Approximation 6 (Assumed density prediction)**.** *The approximation is given by* $p_F(t,x) \approx \mathcal{N}\big(x;\mu^F(t),\Sigma^F(t)\big)$*, where*

$$\begin{aligned}\dot{\mu}^F(t) &= \hat{\mathbb{E}}_t^F\big[a(t,X(t))\big], \\ \dot{\Sigma}^F(t) &= \hat{\mathbb{E}}_t^F\big[a(t,X(t))(X(t)-\mu^F(t))^\mathsf{T}\big] + \hat{\mathbb{E}}_t^F\big[(X(t)-\mu^F(t))a(t,X(t))^\mathsf{T}\big] \\ &\quad + \hat{\mathbb{E}}_t^F\big[Q(t,X(t))\big],\end{aligned}$$

and $\hat{\mathbb{E}}_t^F$ *is the expectation with respect to* $X(t) \sim \mathcal{N}\big(\mu^F(t),\Sigma^F(t)\big)$.

The assumed density prediction is then followed by its update, which is in fact Approximation 1 from Section 2.1.1. For the reader's convenience, it is re-stated in Approximation 7 below.

---

[2] These relations can also be obtained by manipulating the Fokker–Planck equation directly.

Approximation 7 (Assumed density update)**.** *The approximation is given by* $p_F(t_n,x) \approx \mathcal{N}\left(x; \mu^F(t_n), \Sigma^F(t_n)\right)$*, where*

$$\begin{aligned}
S(t_n) &= \hat{\mathbb{V}}^F_{t_n^-}\left[c(t_n, X(t_n))\right] + \hat{\mathbb{E}}^F_{t_n^-}\left[R(t_n, X(t_n))\right],\\
K(t_n) &= \hat{\mathbb{C}}^F_{t_n^-}\left[X(t_n), c(t_n, X(t_n))\right] S^{-1}(t_n),\\
\mu^F(t_n) &= \mu^F(t_n^-) + K(t_n)\Big(y(t_n) - \hat{\mathbb{E}}^F_{t_n^-}\left[c(t_n, X(t_n))\right]\Big),\\
\Sigma^F(t_n) &= \Sigma^F(t_n^-) - K(t_n) S(t_n) K^{\mathsf{T}}(t_n).
\end{aligned}$$

and $\hat{\mathbb{E}}^F_{t_n^-}$*,* $\hat{\mathbb{C}}^F_{t_n^-}$*, and* $\hat{\mathbb{V}}^F_{t_n^-}$ *are expectations, cross-covariances, and covariances with respect to* $X(t_n) \sim \mathcal{N}\left(\mu^F(t_n^-), \Sigma^F(t_n^-)\right)$*, respectively.*

Approximations 6 and 7 together define the continuous-discrete assumed Gaussian density filter. There are various approaches to using the assumed density filter to define an assumed density smoother. One approach is to approximate the operator $\mathscr{K}_1$ by replacing the exact filtering density in Equation (4.6) with the approximation from the assumed density filter, which gives approximate expressions for $\partial_t \mathbb{E}[X(t) \mid \mathscr{Y}(t_N)]$ and $\partial_t \mathbb{V}[X(t) \mid \mathscr{Y}(t_N)]$. Then these expectations are replaced with approximations from the assumed density smoother. The resulting approximation is given in Approximation 8, for which a detailed derivation is given by Särkkä and Sarmavuori (2013).

Approximation 8 (Assumed density smoother I)**.** *The approximation is given by* $p_S(t,x) \approx \mathcal{N}(x; \mu^S(t), \Sigma^S(t))$*, where*

$$\begin{aligned}
\dot{\mu}^S(t) = {}& \hat{\mathbb{E}}^S_t\left[a(t, X(t))\right] - \hat{\mathbb{E}}^S_t\left[Q(t, X(t))[\Sigma^S(t)]^{-1}\left(X(t) - \mu^S(t)\right)\right]\\
&+ \hat{\mathbb{E}}^S_t\left[Q(t, X)[\Sigma^F(t)]^{-1}\left(X(t) - \mu^F(t)\right)\right],\\
\dot{\Sigma}^S(t) = {}& \hat{\mathbb{E}}^S_t\left[a(t, X(t))\left(X(t) - \mu^F(t)\right)^{\mathsf{T}}\right] + \hat{\mathbb{E}}^S_t\left[a(t, X(t))\left(X(t) - \mu^F(t)\right)^{\mathsf{T}}\right]^{\mathsf{T}}\\
&+ \hat{\mathbb{E}}^S_t\left[Q(t, X(t))[\Sigma^F(t)]^{-1}\left(X(t) - \mu^F(t)\right)\left(X(t) - \mu^S(t)\right)^{\mathsf{T}}\right]\\
&+ \hat{\mathbb{E}}^S_t\left[Q(t, X(t))[\Sigma^F(t)]^{-1}\left(X(t) - \mu^F(t)\right)\left(X(t) - \mu^S(t)\right)^{\mathsf{T}}\right]^{\mathsf{T}}\\
&- \hat{\mathbb{E}}^S_t\left[Q(t, X(t))[\Sigma^S(t)]^{-1}\left(X(t) - \mu^S(t)\right)\left(X(t) - \mu^S(t)\right)^{\mathsf{T}}\right]\\
&- \hat{\mathbb{E}}^S_t\left[Q(t, X(t))[\Sigma^S(t)]^{-1}\left(X(t) - \mu^S(t)\right)\left(X(t) - \mu^S(t)\right)^{\mathsf{T}}\right]^{\mathsf{T}}\\
&+ \hat{\mathbb{E}}^S_t\left[Q(t, X(t))\right],
\end{aligned}$$

and $\hat{\mathbb{E}}^S_t$ *is the expectation with respect to* $X(t) \sim \mathcal{N}\left(\mu^S(t), \Sigma^S(t)\right)$*. Additionally,* *if $Q(t,x)$ is constant in x, $Q(t,x) = Q(t)$ then this simplifies to*

$$\dot{\mu}^S(t) = \hat{\mathbb{E}}_t^S\left[a(t,X(t))\right] + Q(t)[\Sigma^F(t)]^{-1}\left(\mu^S(t) - \mu^F(t)\right),$$

$$\dot{\Sigma}^S(t) = \hat{\mathbb{E}}_t^S\left[a(t,X(t))(X(t)-\mu^F(t))^\mathsf{T}\right] + \hat{\mathbb{E}}_t^S\left[a(t,X(t))(X(t)-\mu^F(t))^\mathsf{T}\right]^\mathsf{T}$$
$$+ Q(t)[\Sigma^F(t)]^{-1}\Sigma^S(t) + \left[Q(t)[\Sigma^F(t)]^{-1}\Sigma^S(t)\right]^\mathsf{T} - Q(t).$$

Another approach to deriving assumed density smoothers is to discretise the stochastic differential equation in Equation (3.15) using the Euler–Maruyama scheme, applying the discrete time assumed density smoother (See Section 4.2.2) and take the formal limit as the discretisation interval tends to zero (Särkkä and Sarmavuori 2013, see also Särkkä 2008). This results in Approximation 9.

Approximation 9 (Assumed density smoother II)**.** *The approximation is given by $p_S(t,x) \approx \mathcal{N}\left(x;\mu^S(t),\Sigma^S(t)\right)$, where*

$$\dot{\mu}^S(t) = \hat{\mathbb{E}}_t^S\left[a(t,X(t))\right] + \hat{\mathbb{E}}_t^S[Q(t,X(t))][\Sigma^F(t)]^{-1}\left(\mu^S(t) - \mu^F(t)\right)$$
$$+ \hat{\mathbb{E}}_t^S\left[a(t,X(t))(X(t)-\mu^F(t))^\mathsf{T}\right](\mu^S(t) - \mu^F(t)),$$

$$\dot{\Sigma}^S(t) = \left(\hat{\mathbb{E}}_t^S\left[a(t,X(t))\left(X(t)-\mu^F(t)\right)^\mathsf{T}\right] + \hat{\mathbb{E}}_t^S[Q(t,X(t))]\right)[\Sigma^F(t)]^{-1}\Sigma^S(t)$$
$$+ \left[\left(\hat{\mathbb{E}}_t^S\left[a(t,X(t))\left(X(t)-\mu^F(t)\right)^\mathsf{T}\right] + \hat{\mathbb{E}}_t^S[Q(t,X(t))]\right)[\Sigma^F(t)]^{-1}\Sigma^S(t)\right]^\mathsf{T}$$
$$- \hat{\mathbb{E}}_t^S[Q(t,X(t))],$$

and $\hat{\mathbb{E}}_t^S$ is the expectation with respect to $X(t) \sim \mathcal{N}(\mu^S(t),\Sigma^S(t))$.

Projection Filtering and Smoothing

Similarly to the assumed density approach, the projection method constrains the approximate filtering and smoothing densities to be in a particular class of densities. From the formal prediction relation in Equation (4.4), it follows that the square root of the density evolves as

$$\partial_t p_F^{1/2}(t,x) = \sqrt{\mathscr{A}}^\star p_F^{1/2}(t,x),$$

where the operator $\sqrt{\mathscr{A}}^\star$ acts on $p^{1/2}$ according to

$$\sqrt{\mathscr{A}}^\star p^{1/2} = \frac{p^{-1/2}}{2}\mathscr{A}^\star p.$$

Assuming that $p_F(t_n,x)$ is in some parametric class of densities, $p_F(t_n,x) \in \mathscr{P} = \left\{p_\theta : \theta \in \Theta\right\}$, the projection approach is to use the projection formula in Section 2.2 according to (Brigo et al., 1999, cf. Eq. (15))

$$\partial_t \hat{p}_{\theta_F(t)}^{1/2}(t,x) = \Pi_{\theta_F(t)} \circ \sqrt{\mathscr{A}}^\star \hat{p}_{\theta_F(t)}^{1/2}(t,x). \tag{4.21}$$

Note that $\hat{p}^{1/2}_{\theta_F(t)}(t_n,x) \in \mathscr{P}$ implies that $\hat{p}^{1/2}_{\theta_F(t)}(t,x) \in \mathscr{P}$ for $t \in [t_n, t_{n+1})$. Therefore, Equation (4.21) evolves in a finite dimensional manifold and can, as the notation suggests, be identified with a curve $\theta_F(t)$ in Θ (Brigo et al., 1999). When $\mathscr{P}$ is the class of Gaussian densities, $\theta_F(t) = \big(\mu^F(t), \Sigma^F(t)\big)$, then the projection prediction in Equation (4.21) is the same as the assumed density prediction in Approximation 6 (Koyama, 2018, Eqs. (28) and (29)).

As for the filter update, the projection framework classically does not offer a way to perform Bayes' update unless the class of densities $\mathscr{P}$ is conjugate to the measurement likelihood. One approach to approximate updates with the projection method was proposed in Publication IV (Tronarp and Särkkä, 2019b), this will be discussed in more detail in Chapter 5.

The projection approach to smoothing is completely analogous to that of the prediction. That is, a square root form of the smoothing density can be obtained from Equation (4.7), which is given by

$$\partial_t p_S^{1/2}(t,x) = \sqrt{\mathscr{K}_1}^{\star} p_S^{1/2}(t,x),$$

where the operator $\sqrt{\mathscr{K}_1}^{\star}$ acts on $p^{1/2}$ according to

$$\sqrt{\mathscr{K}_1}^{\star} p^{1/2} = \frac{p^{-1/2}}{2} \mathscr{K}_1^{\star} p.$$

By using the projection formula again, the following relation is obtained for the approximate smoothing density (Koyama, 2018)[3]:

$$\partial_t \hat{p}^{1/2}_{\theta_S(t)}(t,x) = \Pi_{\theta_S(t)} \circ \sqrt{\mathscr{K}_1}^{\star} \hat{p}^{1/2}_{\theta_S(t)}(t,x). \tag{4.22}$$

As it is not obvious from the notation, it should be noted that the output of the projection filter $\hat{p}_{\theta_F(t)}$ replaces the exact filtering distribution p_F in the expression for $\sqrt{\mathscr{K}_1}^{\star}$ in Equation (4.22) (recall the definition of $\mathscr{K}_1^{\star}$ in Equation (4.8)). Furthermore, if $\mathscr{P}$ is the class of Gaussian densities, $\theta_F(t) = \big(\mu^F(t), \Sigma^F(t)\big)$ and $\theta_S(t) = \big(\mu^S(t), \Sigma^S(t)\big)$, then the projection smoother in Equation (4.22) reduces to the assumed density smoother in Approximation 8 (Koyama, 2018, Equations (30) and (31)).

4.2 Bayesian Inference in Discrete-Time Models

In this section, inference in discrete time models is considered. The formal solution to the filtering and smoothing problems is reviewed. Both problems are considerably simpler than in the continuous-discrete setting, in the sense that their solutions only involve integrals rather than partial differential equations (Särkkä, 2013). Following this, inference in affine Gaussian models is discussed, which can be solved by the discrete-time

---

[3] Koyama (2018) uses the expression obtained by Anderson (1972) for the formal smoothing solution.

versions of the Kalman filter (Kalman, 1960) and the Rauch–Tung–Striebel smoother (Rauch et al., 1965). Then, the assumed density method in discrete time is reviewed as well as its iterated variants (Bell, 1994, Bell and Cathey, 1993, García-Fernández et al., 2014, 2015, 2017). Lastly, a brief overview of sequential Monte Carlo methods is given.

4.2.1 The Formal Solution

The formal solution of the discrete time inference problem involves finding the filtering and smoothing densities of the following state-space model:

$$\begin{aligned} X_0 &\sim p_{X_0}(x), \\ X_n \mid X_{n-1} &\sim p_{X_n \mid X_{n-1}}(x \mid x_{n-1}), \\ Y_n \mid X_n &\sim p_{Y_n \mid X_n}(y \mid x_n). \end{aligned}$$

As with the continuous-time inference problem this entails that exact mathematical relations between filtering densities, likelihoods, and smoothing densities are given.

The Formal Filtering Relations
The filtering density is denoted by $p_F(n-1,x) = p\big(n-1,x \mid \mathscr{Y}(n-1)\big)$. As the dynamics is defined by a Markov model, it follows that the one step-ahead prediction density is given by (Särkkä, 2013, Theorem 4.1)

$$p_F^-(n,x) = \int_{\mathbb{X}} p_{X_n \mid X_{n-1}}(x \mid u) p_F(n-1,u)\,\mathrm{d}u. \tag{4.24}$$

The filter update is the same as for the continuous-discrete case. That is,

$$p_F(x,n) = \frac{L_n(x) p_F^-(n,x)}{\int_{\mathbb{X}} L_n(x) p_F^-(n,x)\,\mathrm{d}x}. \tag{4.25}$$

The Formal Smoothing Relations
Denote the smoothing density by $p_S(n,x) = p\big(n,x \mid \mathscr{Y}(N)\big)$. Then the smoothing recursion is given by (Särkkä, 2013, Theorem 8.1)

$$p_S(n,x) = p_F(n,x) \int_{\mathbb{X}} \frac{p_{X_{n+1} \mid X_n}(u \mid x)}{p_F^-(n+1,u)} p_S(n+1,u)\,\mathrm{d}u. \tag{4.26}$$

4.2.2 Inference in Affine Models

As with the continuous-discrete time case, ”closed-form“ solutions to the discrete-time inference problem are elusive. However, for affine models there are discrete-time versions of the Kalman filter (Kalman, 1960) and Rauch–Tung–Striebel smoother (Rauch et al., 1965). These recursive algorithms are described in the sequel.

Discrete-Time Kalman Filter

Just as in continuous-discrete time, the discrete time Kalman filter operates by alternating between predictions and updates. If the filtering density at time $n-1$ is $p_F(n-1,x) = \mathcal{N}\left(x;\mu_{n-1}^F,\Sigma_{n-1}^F\right)$ then the prediction density at time n is given by $p_F^-(n,x) = \mathcal{N}\left(x;\mu_n^P,\Sigma_n^P\right)$ where

$$\begin{aligned}\mu_n^P &= A_n\mu_{n-1}^F + b_n,\\ \Sigma_n^P &= A_n\Sigma_{n-1}^F A_n^\top + Q_n.\end{aligned}$$

Furthermore, the filtering density at time n is $p_F(n,x) = \mathcal{N}\left(x;\mu_n^F,\Sigma_n^F\right)$ and the parameters are given by

$$\begin{aligned}S_n &= C_n\Sigma_n^P C_n^\top + R_n,\\ K_n &= \Sigma_n^P C_n^\top S_n^{-1},\\ \mu_n^F &= \mu_{n-1}^P + K_n\left(y_n - C_n\mu_n^P - d_n\right),\\ \Sigma_n^F &= \Sigma_n^P - K_nS_nK_n^\top,\end{aligned}$$

which follows directly from Lemma 1.

Discrete-Time Rauch–Tung–Striebel Smoother

The smoothing density at the terminal time stamp N is given by $p_S(N,x) = p_F(N,x)$ and for $n \leq N$ it is given by $p_S(n,x) = \mathcal{N}\left(x;\mu_n^S,\Sigma_n^S\right)$. The discrete-time Rauch–Tung–Striebel smoother is then a backwards recursion for the parameters μ_n^S and Σ_n^S, which is given by (Rauch et al., 1965)

$$G_n = \Sigma_n^F A_{n+1}^\top\left[\Sigma_{n+1}^P\right]^{-1}, \tag{4.29a}$$

$$\mu_n^S = \mu_n^F + G_n\left(\mu_{n+1}^S - \mu_{n+1}^P\right), \tag{4.29b}$$

$$\Sigma_n^S = \Sigma_n^F + G_n\left(\Sigma_{n+1}^S - \Sigma_{n+1}^P\right)G_n^\top. \tag{4.29c}$$

4.2.3 Inference in Non-Affine Models

Just as in continuous-discrete time, the assumed density approach is a popular tool for the discrete time inference problem. In fact, for the assumed density approach considerable research effort has been put towards approximating the necessary expectations (Arasaratnam and Haykin, 2009, Arasaratnam et al., 2007, Ito and Xiong, 2000, Julier et al., 2000, Prüher and Straka, 2017, Prüher and Šimandl, 2015, Särkkä, 2008, Wu et al., 2006) and formulating iterative variants (Bell, 1994, Bell and Cathey, 1993, García-Fernández et al., 2014, 2015, 2017, Skoglund et al., 2015, Tronarp et al., 2018a). Another popular approach is sequential Monte Carlo (Arulampalam et al., 2002, Cappé et al., 2007, Del Moral et al., 2006, Doucet et al., 2000, Godsill et al., 2004, Gordon et al., 1993). In the following the assumed Gaussian density approach and the particle filtering approaches are reviewed.

The Assumed Density Approach

In the discrete time, the assumed density approach approximates the left-hand sides of the formal filtering and smoothing solutions Equations (4.24), (4.25), and (4.26) is approximated within a specific class of densities. Although the assumed density is most commonly taken to be Gaussian, quite a few assumed density approaches based on other classes of densities have cropped up over the years. For instance, there are approaches based on the Student's t distribution (Huang et al., 2016, Prüher et al., 2017, Roth et al., 2013, Tronarp et al., 2016), the von Mises–Fisher distribution (Bukal et al., 2017, Kurz et al., 2016, Marković et al., 2014a,b, Traa and Smaragdis, 2014), and the Bingham distribution (Gilitschenski et al., 2014, 2016, Glover and Kaelbling, 2014, Kurz et al., 2013, 2014). The exposition here is restricted to the assumed Gaussian approach for conditionally Gaussian systems. The following equations for the predictive mean and covariance follow from the Markov property

$$\mathbb{E}\left[X_n \mid \mathscr{Y}_{n-1}\right] = \mathbb{E}\left[a_n(X_{n-1}) \mid \mathscr{Y}_{n-1}\right], \tag{4.30a}$$

$$\mathbb{V}\left[X_n \mid \mathscr{Y}_{n-1}\right] = \mathbb{V}\left[a_n(X_{n-1}) \mid \mathscr{Y}_{n-1}\right] + \mathbb{E}\left[Q_n(X_{n-1}) \mid \mathscr{Y}_{n-1}\right], \tag{4.30b}$$

see Equation (3.26). Replacing the expectations in Equation (4.30) with expectations with respect to the Gaussian approximation to the filtering density gives the assumed Gaussian prediction equations, which are given in Approximation 10.

Approximation 10 (Assumed density prediction). *The approximation is given by* $p_F^-(n,x) \approx \mathcal{N}\left(x; \mu_n^P, \Sigma_n^P\right)$*, where*

$$\begin{aligned}
\mu_n^P &= \hat{\mathbb{E}}_{n-1}^F\left[a_n(X_{n-1})\right],\\
\Sigma_n^P &= \hat{\mathbb{V}}_{n-1}^F\left[a_n(X_{n-1})\right] + \hat{\mathbb{E}}_{n-1}^F\left[Q_n(X_{n-1})\right],
\end{aligned}$$

and $\hat{\mathbb{E}}_{n-1}^F$ *and* $\hat{\mathbb{V}}_{n-1}^F$ *are expectations and covariances with respect to* $X_{n-1} \sim \mathcal{N}\left(\mu_{n-1}^F, \Sigma_{n-1}^F\right)$*, respectively.*

Since the exact update in discrete time is structurally the same as in continuous-discrete time, it should come as no surprise that the assumed Gaussian update equations are the same as well, which are re-stated in Approximation 11.

Approximation 11 (Assumed density update). *The approximation is given by* $p_F(n,x) \approx \mathcal{N}\left(x; \mu_n^F, \Sigma_n^F\right)$*, where*

$$\begin{aligned}
S_n &= \hat{\mathbb{V}}_n^P\left[c_n(X_n)\right] + \hat{\mathbb{E}}_n^P\left[R_n(X_n)\right],\\
K_n &= \hat{\mathbb{C}}_n^P\left[X_n, c_n(X_n)\right] S_n^{-1},\\
\mu_n^F &= \mu_n^P + K_n\left(y_n - \hat{\mathbb{E}}_n^P[c_n(X_n)]\right),\\
\Sigma_n^F &= \Sigma_n^P - K_n S_n K_n^\mathsf{T},
\end{aligned}$$

and $\hat{\mathbb{E}}_n^P$, $\hat{\mathbb{C}}_n^P$, *and* $\hat{\mathbb{V}}_n^P$ *are expectations, cross-covariances, and covariances with respect to* $X_n \sim \mathcal{N}\big(\mu_n^P, \Sigma_n^P\big)$, *respectively.*

Finally, the assumed Gaussian smoother follows by noticing that G_n in Equation (4.29) in moment form is given by

$$G_n = \mathbb{C}\big[X_n, X_{n+1} \mid \mathcal{Y}_n\big] \mathbb{V}\big[X_{n+1} \mid \mathcal{Y}_n\big]^{-1}.$$

Replacing these moments with the moments with respect to the Gaussian approximation gives the assumed Gaussian smoother in Approximation 12 (Särkkä and Hartikainen, 2010).

Approximation 12 (Assumed density smoother)**.** *The approximation is given by* $p_S(n,x) \approx \mathcal{N}\big(x; \mu_n^S, \Sigma_n^S\big)$, *where*

$$\begin{aligned}
G_n &= \hat{\mathbb{C}}_n^F \big[X_n, a_{n+1}(X_{n+1})\big] \big[\Sigma_{n+1}^P\big]^{-1}, \\
\mu_n^S &= \mu_n^F + G_n \big(\mu_{n+1}^S - \mu_{n+1}^P\big), \\
\Sigma_n^S &= \Sigma_n^F + G_n \big(\Sigma_{n+1}^S - \Sigma_{n+1}^P\big) G_n^\top,
\end{aligned}$$

and $\hat{\mathbb{C}}_n^F$ *is the cross-covariance with respect to* $X_n \sim \mathcal{N}\big(\mu_n^F, \Sigma_n^F\big)$.

Iterated Posterior Linearisation

In order to understand the iterated posterior linearisation method, it is instructive to re-write the assumed density method, Approximations 10, 11, and 12, in statistical linear regression form, which are given by Approximations 13, 14, and 15.

Approximation 13 (Statistical linear regression prediction)**.** *The approximation is given by* $p_F^-(n,x) \approx \mathcal{N}\big(x; \mu_n^P, \Sigma_n^P\big)$, *where*

$$\begin{aligned}
\mu_n^P &= \hat{A}_n \mu_{n-1}^F + \hat{b}_n, \\
\Sigma_n^P &= \hat{A}_n \Sigma_{n-1}^F \hat{A}_n^\top + \hat{Q}_n,
\end{aligned}$$

with

$$\hat{A}_n = \hat{\mathbb{C}}_{n-1}^F \big[a_n(X_{n-1}), X_{n-1}\big] \big[\Sigma_{n-1}^F\big]^{-1}, \tag{4.35a}$$

$$\hat{b}_n = \hat{\mathbb{E}}_{n-1}^F \big[a_n(X_{n-1})\big] - \hat{A}_n \mu_{n-1}^F, \tag{4.35b}$$

$$\hat{Q}_n = \hat{\mathbb{V}}_{n-1}^F \big[a_n(X_{n-1})\big] + \hat{\mathbb{E}}_{n-1}^F \big[Q_n(X_{n-1})\big] - \hat{A}_n \Sigma_{n-1}^F \hat{A}_n^\top, \tag{4.35c}$$

and $\hat{\mathbb{E}}_{n-1}^F$, $\hat{\mathbb{C}}_{n-1}^F$, *and* $\hat{\mathbb{V}}_{n-1}^F$ *are expectations, cross-covariances, and covariances with respect to* $X_{n-1} \sim \mathcal{N}\big(\mu_{n-1}^F, \Sigma_{n-1}^F\big)$, *respectively.*

Approximation 14 (Statistical linear regression update)**.** *The approximation is given by* $p_F(n,x) \approx \mathcal{N}(x;\mu_n^F,\Sigma_n^F)$*, where*

$$S_n = \hat{C}_n \Sigma_n^P \hat{C}_n^\top + \hat{R}_n, \tag{4.36a}$$

$$K_n = \Sigma_n^P \hat{C}_n^\top S_n^{-1}, \tag{4.36b}$$

$$\mu_n^F = \mu_n^P + K_n\left(y_n - \hat{C}_n \mu_n^P - \hat{d}_n\right), \tag{4.36c}$$

$$\Sigma_n^F = \Sigma_n^P - K_n S_n K_n^\top, \tag{4.36d}$$

with

$$\hat{C}_n = \hat{\mathbb{C}}_n^P\left[c_n(X_n),X_n\right]\left[\Sigma_n^P\right]^{-1}, \tag{4.37a}$$

$$\hat{d}_n = \hat{\mathbb{E}}_n^P\left[c_n(X_n)\right] - \hat{C}_n \mu_n^P, \tag{4.37b}$$

$$\hat{R}_n = \hat{\mathbb{V}}_n^P\left[c_n(X_n)\right] + \hat{\mathbb{E}}_n^P\left[R_n(X_n)\right] - \hat{C}_n \Sigma_n^P \hat{C}_n^\top, \tag{4.37c}$$

and $\hat{\mathbb{E}}_n^P$*,* $\hat{\mathbb{C}}_n^P$*, and* $\hat{\mathbb{V}}_n^P$ *are expectations, cross-covariances, and covariances with respect to* $X_n \sim \mathcal{N}\left(\mu_n^P,\Sigma_n^P\right)$*, respectively.*

Approximation 15 (Statistical linear regression smoother)**.** *The approximation is given by* $p_S(n,x) \approx \mathcal{N}\left(x;\mu_n^S,\Sigma_n^S\right)$*, where*

$$\begin{aligned}
G_n &= \Sigma_n^F \hat{A}_{n+1}^\top \left[\Sigma_{n+1}^P\right]^{-1}, \\
\mu_n^S &= \mu_n^F + G_n\left(\mu_{n+1}^S - \mu_{n+1}^P\right), \\
\Sigma_n^S &= \Sigma_n^F + G_n\left(\Sigma_{n+1}^S - \Sigma_{n+1}^P\right) G_n^\top.
\end{aligned}$$

The statistical linear regression filter and smoother are generalisations of the extended Kalman filter and smoother, respectively. In fact if the first order Taylor series method (Approximation 4) is used to approximate the linearisation parameters the following is obtained for the prediction

$$\begin{aligned}
\hat{A}_n &\approx J_{a_n}\left(\mu_{n-1}^F\right), \\
\hat{b}_n &= a_n\left(\mu_{n-1}^F\right) - J_{a_n}\left(\mu_{n-1}^F\right)\mu_{n-1}^F, \\
\hat{Q}_n &= Q_n\left(\mu_{n-1}^F\right),
\end{aligned}$$

and

$$\begin{aligned}
\mu_n^P &= a_n\left(\mu_{n-1}^F\right), \\
\Sigma_n^P &= J_{a_n}\left(\mu_{n-1}^F\right)\Sigma_{n-1}^F J_{a_n}^\top\left(\mu_{n-1}^F\right) + Q_n\left(\mu_{n-1}^F\right),
\end{aligned}$$

which is the prediction of the standard extended Kalman filter when Q_n does not depend on the state (Särkkä, 2013). Similarly for the filter update the following holds when using Approximation 4

$$\begin{aligned}
\hat{C}_n &\approx J_{c_n}\left(\mu_n^P\right), \\
\hat{d}_n &= c_n\left(\mu_n^P\right) - J_{c_n}\left(\mu_n^P\right)\mu_n^P, \\
\hat{R}_n &= R_n\left(\mu_n^P\right),
\end{aligned}$$

$$\begin{aligned}
S_n &= J_{c_n}\big(\mu_n^P\big)\Sigma_n^P J_{c_n}^\top\big(\mu_n^P\big) + R_n\big(\mu_n^P\big),\\
K_n &= \Sigma_n^P J_{c_n}^\top\big(\mu_n^P\big) S_n^{-1},\\
\mu_n^F &= \mu_n^P + K_n\big(y_n - c_n\big(\mu_n^P\big)\big),\\
\Sigma_n^F &= \Sigma_n^P - K_n S_n K_n^\top,
\end{aligned}$$

which is the filter update of the standard extended Kalman filter when Q_n does not depend on the state (Särkkä, 2013). Lastly, the smoother gain is now given by

$$G_n = \Sigma_n^F J_{a_{n+1}}^\top\big(\mu_n^F\big)\big[\Sigma_{n+1}^P\big]^{-1},$$

which is precisely how the smoother gain as computed by the extended Kalman smoother (Särkkä, 2013).

The similarities between statistical linear regression and extened Kalman filtering/smoothing can also be understood by using integration by parts, which gives $\hat{A}_n$ and $\hat{C}_n$ as

$$\begin{aligned}
\hat{A}_n &= \mathbb{E}_{n-1}^F\big[J_{a_n}\big(X_{n-1}\big)\big],\\
\hat{C}_n &= \mathbb{E}_n^P\big[J_{c_n}\big(X_n\big)\big].
\end{aligned}$$

Consequently, the prediction is given by

$$\begin{aligned}
\mu_n^P &= \mathbb{E}_{n-1}^F\big[a_n\big(X_{n-1}\big)\big],\\
\Sigma_n^P &= \mathbb{E}_{n-1}^F\big[J_{a_n}\big(X_{n-1}\big)\big]\Sigma_{n-1}^F\mathbb{E}_{n-1}^F\big[J_{a_n}\big(X_{n-1}\big)\big]^\top + \hat{Q}_n,
\end{aligned}$$

the update is given by

$$\begin{aligned}
S_n &= \mathbb{E}_n^P\big[J_{c_n}\big(X_n\big)\big]\Sigma_n^P\mathbb{E}_n^P\big[J_{c_n}\big(X_n\big)\big]^\top + \hat{R}_n,\\
K_n &= \Sigma_n^P\mathbb{E}_n^P\big[J_{c_n}\big(X_n\big)\big]^\top S_n^{-1},\\
\mu_n^F &= \mu_n^P + K_n\big(y_n - \mathbb{E}_n^P\big[c_n\big(X_n\big)\big]\big),\\
\Sigma_n^F &= \Sigma_n^P - K_n S_n K_n^\top,
\end{aligned}$$

and the smoother gain is given by

$$G_n = \Sigma_n^F\mathbb{E}_n^F\big[J_{a_{n+1}}\big(X_n\big)\big]^\top\big[\Sigma_{n+1}^P\big]^{-1}.$$

That is, rather than using point evaluations the statistical linear regression method is using averaging.

The iterated posterior linearisation method for smoothing proceeds iteratively by replacing the expectations, cross-covariances, and covariances in Equations (4.35) and (4.37) with the current best approximation to the corresponding smoothing marginals (García-Fernández et al., 2017, Tronarp et al., 2018a). More specifically, denote the approximate filter, prediction, and smoother marginals at iteration l by $p_{F,l}(n,x) = \mathcal{N}\left(x; \mu^F_{n,l}, \Sigma^F_{n,l}\right)$, $p^-_{F,l}(n,x) = \mathcal{N}\left(x; \mu^P_{n,l}, \Sigma^P_{n,l}\right)$, and $p_{S,l}(n,x) = \mathcal{N}\left(x; \mu^S_{n,l}, \Sigma^S_{n,l}\right)$, $n = 0,\ldots,N$, respectively. Then the approximate filtering, prediction, and smoothing marginals at iteration $l+1$ are obtained through Approximations 16, 17, 18.

Remark 8. *The posterior linearisation approach to filtering simply alternates between the standard assumed density prediction, Approximation 10, and the iterative update described in Approximation 3.*

Approximation 16 (Iterative statistical linear regression prediction)**.** *The approximation is given by* $p^-_{F,l+1}(n,x) \approx \mathcal{N}\left(x; \mu^P_{n,l+1}, \Sigma^P_{n,l+1}\right)$*, where*

$$\begin{aligned}
\mu^P_{n,l+1} &= \hat{A}_{n,l}\mu^F_{n-1,l+1} + \hat{b}_{n,l},\\
\Sigma^P_{n,l+1} &= \hat{A}_{n,l}\Sigma^F_{n-1,l+1}\hat{A}^\top_{n,l} + \hat{Q}_{n,l},
\end{aligned}$$

with

$$\begin{aligned}
\hat{A}_{n,l} &= \hat{\mathbb{C}}^S_{n-1,l}\left[a_n(X_{n-1}), X_{n-1}\right]\left[\Sigma^S_{n-1,l}\right]^{-1},\\
\hat{b}_{n,l} &= \hat{\mathbb{E}}^S_{n-1,l}\left[a_n(X_{n-1})\right] - \hat{A}_{n,l}\mu^S_{n-1,l},\\
\hat{Q}_{n,l} &= \hat{\mathbb{V}}^S_{n-1,l}\left[a_n(X_{n-1})\right] + \hat{\mathbb{E}}^S_{n-1,l}\left[Q_n(X_{n-1})\right] - \hat{A}_{n,l}\Sigma^S_{n-1,l}\hat{A}^\top_{n,l},
\end{aligned}$$

and $\hat{\mathbb{E}}^S_{n-1,l}$, $\hat{\mathbb{C}}^S_{n-1,l}$, *and* $\hat{\mathbb{V}}^S_{n-1,l}$ *are expectations, cross-covariances, and covariances with respect to* $X_{n-1} \sim \mathcal{N}\left(\mu^S_{n-1,l}, \Sigma^S_{n-1,l}\right)$*, respectively.*

Approximation 17 (Iterative statistical linear regression update)**.** *The approximation is given by* $p_{F,l+1}(n,x) \approx \mathcal{N}\left(x; \mu^F_{n,l+1}, \Sigma^F_{n,l+1}\right)$*, where*

$$\begin{aligned}
S_{n,l+1} &= \hat{C}_{n,l}\Sigma^P_{n,l+1}\hat{C}^\top_{n,l} + \hat{R}_{n,l},\\
K_{n,l+1} &= \Sigma^P_{n,l+1}\hat{C}^\top_{n,l}S^{-1}_{n,l+1},\\
\mu^F_{n,l+1} &= \mu^P_{n,l+1} + K_{n,l+1}\left(y_n - \hat{C}_{n,l}\mu^P_{n,l+1} - \hat{d}_{n,l}\right),\\
\Sigma^F_{n,l+1} &= \Sigma^P_{n,l+1} - K_{n,l+1}S_{n,l+1}K^\top_{n,l+1}
\end{aligned}$$

with

$$\begin{aligned}
\hat{C}_{n,l} &= \hat{\mathbb{C}}^S_{n,l}\left[c_n(X_n), X_n\right]\left[\Sigma^S_{n,l}\right]^{-1},\\
\hat{d}_{n,l} &= \hat{\mathbb{E}}^S_{n,l}\left[c_n(X_n)\right] - \hat{C}_{n,l}\mu^S_{n,l},\\
\hat{R}_{n,l} &= \hat{\mathbb{V}}^S_{n,l}\left[c_n(X_n)\right] + \hat{\mathbb{E}}^S_{n,l}\left[R_n(X_n)\right] - \hat{C}_n\Sigma^P_n\hat{C}^\top_n,
\end{aligned}$$

and $\hat{\mathbb{E}}^S_{n,l}$, $\hat{\mathbb{C}}^S_{n,l}$, *and* $\hat{\mathbb{V}}^S_{n,l}$ *are expectations, cross-covariances, and covariances with respect to* $X_n \sim \mathcal{N}\left(\mu^S_{n,l}, \Sigma^S_{n,l}\right)$*, respectively.*

Approximation 18 (Statistical linear regression smoother)**.** *The approximation is given by* $p_S(n,x) \approx \mathcal{N}\big(x;\mu^S_{n,l+1},\Sigma^S_{n,l+1}\big)$*, where*

$$\begin{aligned}
G_{n,l+1} &= \Sigma^F_{n,l+1}\hat{A}^\top_{n+1,l}\left[\Sigma^P_{n+1,l+1}\right]^{-1},\\
\mu^S_{n,l+1} &= \mu^F_{n,l+1} + G_{n,l+1}\big(\mu^S_{n+1,l+1} - \mu^P_{n+1,l+1}\big),\\
\Sigma^S_{n,l+1} &= \Sigma^F_{n,l+1} + G_{n,l+1}\big(\Sigma^S_{n+1,l+1} - \Sigma^P_{n+1,l+1}\big)G^\top_{n,l+1}.
\end{aligned}$$

Remark 9. *In the original derivation, the posterior linearisation method (García-Fernández et al., 2014, 2015, 2017) assumed* $Q_n(x) = Q_n$ *and* $R_n(x) = R_n$*. The extension to x dependent* Q_n *and* R_n *is part of the contribution of Publication II.*

Remark 10. *If* $Q_n(x) = Q_n$*,* $R_n(x) = R_n$*, and the expectations in Approximations 16, 17, and 18 are approximated with Taylor series expansions, then the method reduces to the iterated extended Kalman smoother (Bell, 1994, Bell and Cathey, 1993). This follows immediately from the previous discussion of the connections between statistical linear regression filtering and smoothing with extended Kalman filtering and smoothing.*

Particle Filtering

In quite a significant departure of the above approaches, the particle filter is based on importance sampling rather than approximating the filtering density in a fixed class of densities (Särkkä, 2013, Chapter 7). That is, for some importance density $q_{X_{0:n}}(x_{0:n})$ and function $\phi(x_{0:n})$ the expectation of $\phi(X_{0:n})$ given the data $\mathscr{Y}_n$ can be written as (Cappé et al., 2007)

$$\begin{aligned}
\mathbb{E}\big[\phi(X_{0:n}) \mid \mathscr{Y}_n\big] &= \int_{\mathbb{X}} \phi(x_{0:n}) p_{X_{0:n}|Y_{1:n}}(x_{0:n} \mid y_{1:n})\,\mathrm{d}x_{0:n}\\
&= \int_{\mathbb{X}} \phi(x_{0:n}) W(x_{0:n}) q_{X_{0:n}}(x_{0:n})\,\mathrm{d}x_{0:n}.
\end{aligned} \tag{4.51}$$

where W is the likelihood ratio

$$W(x_{0:n}) = \frac{p_{X_{0:n}|Y_{1:n}}(x_{0:n} \mid y_{1:n})}{q_{X_{0:n}}(x_{0:n})}.$$

This expectation can then be approximated by sampling $X_{0:n,l} \sim q_{X_{0:n}}(x_{0:n})$ and then computing

$$\mathbb{E}\big[\phi(X_{0:n}) \mid \mathscr{Y}_n\big] \approx \hat{\mathbb{E}}\big[\phi(X_{0:n}) \mid \mathscr{Y}_n\big] = \frac{L^{-1}\sum_{l=1}^L \phi(X_{0:n,l})W(X_{0:n,l})}{L^{-1}\sum_{l=1}^L W(X_{0:n,l})}. \tag{4.52}$$

The particle filter is then a clever algorithm to sample $X_{0:n}$ recursively, and circumvent the fact that $p_{X_{0:n}|Y_{1:n}}(X_{0:n,l} \mid y_{1:n})$ can typically only be evaluated up to proportionality[4]. That is, the importance density is constructed to

---

[4] When $p_{X_{0:n}|Y_{1:n}}(X_{0:n,l} \mid y_{1:n})$ can be evaluated exactly the denominator in Equation (4.52) is unnecessary, even foolish.

have the following factorisation[5] (Cappé et al., 2007)

$$q_{X_{0:n}}(x_{0:n}) = q_{X_n|X_{n-1}}(x_n \mid x_{n-1}) q_{X_{0:n-1}}(x_{0:n-1}),$$

which entails the following recursion for the likelihood ratio

$$\begin{aligned} W(x_{0:n}) &\propto \frac{L_n(x_n) p_{X_n|X_{n-1}}(x_n \mid x_{n-1}) p_{X_{0:n-1}|Y_{1:n-1}}(x_{0:n-1} \mid y_{1:n-1})}{q_{X_n|X_{0:n-1}}(x_n \mid x_{0:n-1}) q_{X_{0:n-1}}(x_{0:n-1})} \\ &= \frac{L_n(x_n) p_{X_n|X_{n-1}}(x_n \mid x_{n-1})}{q_{X_n|X_{0:n-1}}(x_n \mid x_{0:n-1})} \frac{p_{X_{0:n}|Y_{1:n}}(x_{0:n} \mid y_{1:n})}{q_{X_{0:n-1}}(x_{0:n-1})} \\ &= \frac{L_n(x_n) p_{X_n|X_{n-1}}(x_n \mid x_{n-1})}{q_{X_n|X_{0:n-1}}(x_n \mid x_{0:n-1})} W(x_{0:n-1}). \end{aligned} \tag{4.53}$$

The recursion for the particle filter is thus as follows. The trajectories $\{X_{0:n-1,l}\}_{l=1}^{L}$ with weights $\{w_{n-1,l}\}_{l=1}^{L}$ have been sampled at time $n-1$. Each trajectory is then advanced to time n by sampling (Arulampalam et al., 2002)

$$X_{n,l} \sim q_{X_n|X_{0:n-1}}(x \mid X_{0:n-1,l}),$$

and the weights are advanced by

$$w_{n,l} = \frac{\frac{L_n(X_{n,l}) p_{X_n|X_{n-1}}(X_{n,l}|X_{n-1,l})}{q_{X_n|X_{0:n-1}}(X_{n,l}|X_{0:n-1,l})} W(X_{0:n-1,l})}{\sum_{l=1}^{L} \frac{L_n(X_{n,l}) p_{X_n|X_{n-1}}(X_{n,l}|X_{n-1,l})}{q_{X_n|X_{0:n-1}}(X_{n,l}|X_{0:n-1,l})} W(X_{0:n-1,l})}, \tag{4.54}$$

where the denominator of Equation (4.54) compensates for the fact that the numerator is proportional rather than equal to the likelihood ratio at time n (see Equation (4.53)). From inspection of Equations (4.51) and (4.52), it would appear the algorithm is done, in fact so far the sequential importance sampling algorithm has been described (Särkkä, 2013, Section 7.3). However, sequential importance sampling suffers from the so-called degeneracy problem (Doucet et al., 2001), which is solved by adding a re-sampling step (see Douc and Cappé 2005 for a review on re-sampling schemes). For instance, multinomial sampling operates by sampling L indices $\{I_l\}_{l=1}^{L}$ that are independently drawn from the categorical distribution over $1,\ldots,L$ with weights $w_1,\ldots,w_L$ and then assigning I_l to the lth trajectory. That is,

$$X_{0:n,l} \leftarrow X_{0:n,I_l},$$

where

$$\mathbb{P}(I_l = k) = w_{n,k}.$$

The performance of the particle filter is heavily dependent on the choice of importance density $q_{X_{0:N}}$, which determines the convergence properties of the algorithm (see, e.g., Crisan and Doucet 2002). In the bootstrap filter, the importance density is selected as $q_{X_n|X_{0:n-1}} = p_{X_n|X_{n-1}}$ (Gordon et al.,

---

[5] $q_{X_{0:n}}$ is allowed to depend on $y_{1:n}$ and $q_{X_n|X_{n-1}}(x_n \mid x_{n-1})$ is allowed to depend on y_n, this is omitted in the notation here.

1993). However, other approaches try to approximate the locally optimal $q_{X_n \mid X_{0:n-1}}$ (Doucet et al., 2000), which is given by

$$q^{\star}_{X_n \mid X_{0:n-1}} \propto L_n(x_n) p_{X_n \mid X_{n-1}}(x_n \mid x_{n-1}). \tag{4.55}$$

While the particle filter targets the smoothing distribution in theory, there are problems in practice (Kitagawa, 1996). Namely, when using re-sampling, it would happen that more and more trajectories share a distant past. Consequently, the algorithm does not sample the state-space efficiently. On the other hand, if re-sampling is not used then the issue of degeneracy rears its ugly head again. Consequently specialised algorithms have been developed to sample efficiently from the smoothing distribution (Godsill et al., 2004), or in other ways post-process the particle filter output (Doucet et al., 2000).

5. Discussion

This chapter provides a discussion on the contributions of the present thesis. The contribution of each publication is described briefly followed by some reflections on the results and future directions of study. Briefly put, in Publication I the discrete-time posterior linearisation method (see Section 4.2.3) is extended to the continuous-discrete-time setting. In Publication II the posterior linearisation method is extended to a more general class of state-space models than it was originally defined for (cf., García-Fernández et al. 2014, 2015, 2017). In Publication III, a contribution to probabilistic numerics is given (Hennig et al., 2015), where the solution to an ordinary differential equation is modelled by a continuous-discrete state-space model for which the inference strategies in Sections 4.1 and 4.2 can be employed for approximate inference, which generalises the approaches of Kersting and Hennig (2016) and Schober et al. (2019). In Publication IV the projection methods of information geometry (see Sections 2.2 and 4.1) are employed to develop an approximate update formula for the Bayesian filter. In Publication V, a model for tracking a norm-constrained vector is developed together with a continuous-discrete assumed density filter based on the von Mises–Fisher distribution. Lastly, in Publication VI basis expansions of the Wiener process (see Section 3.1.1) are employed to develop continuous-discrete assumed Gaussian filters and smoothers, which generalises the result of Lyons et al. (2014).

5.1 Publication I

The aim of this contribution was to define statistical linear regression (see Section 2.1.2) to stochastic differential equations. That is, finding an affine approximation to

$$\mathrm{d}X(t) = a\big(t, X(t)\big)\,\mathrm{d}t + \sigma\big(t, X(t)\big)\,\mathrm{d}W(t),$$

which entail the following approximations of a and σ

$$a(t,x) \approx A(t)x + b(t),$$

$$\sigma(t,x) \approx \bar{\sigma}(t),$$

and the affine approximation is thus given by

$$\mathrm{d}X(t) \approx A(t)X(t)\,\mathrm{d}t + b(t)\,\mathrm{d}t + \bar{\sigma}(t)\,\mathrm{d}W(t).$$

Denote the linearising process by $\{\widehat{X}(t)\}_{0\leq t\leq T}$ and define the auxiliary process

$$\mathrm{d}\widetilde{X}(t) = a\left(t,\widehat{X}(t)\right)\mathrm{d}t + \sigma\left(t,\widehat{X}(t)\right)\mathrm{d}\widetilde{W}(t), \tag{5.2}$$

where $\widetilde{W}$ is a standard Wiener process, which is independent of $\widehat{X}$. An approximation of the auxiliary process $\{\widetilde{X}_a(t)\}_{0\leq t\leq T}$ is then formed, $\widetilde{X}_a(t) \approx \widetilde{X}(t)$ and

$$\mathrm{d}\widetilde{X}_a(t) = A(t)\widetilde{X}_a(t)\,\mathrm{d}t + b(t)\,\mathrm{d}t + \bar{\sigma}(t)\,\mathrm{d}\widehat{W}(t),$$

where $\widehat{W}(t)$ is a standard Wiener process, which is independent of $\widetilde{X}(0)$.

5.1.1 The Main Findings

Two different ways to obtain the parameters $A(t)$, $b(t)$, and $\bar{\sigma}(t)$ were found. The first one is based on the Euler–Maruyama discretisation of Equation (5.2) whereafter the standard statistical linear regression method of Section 2.1.2 is employed locally to obtain $A(t)$, $b(t)$, and $\bar{\sigma}(t)$ at each time point. The parameters are then given by

$$A(t) = \mathbb{C}\left[a(t,\widehat{X}(t)),\widehat{X}(t)\right]\mathbb{V}\left[\widehat{X}(t)\right]^{-1}, \tag{5.3a}$$

$$b(t) = \mathbb{E}\left[a(t,X(t))\right] - A(t)\mathbb{E}\left[\widehat{X}(t)\right], \tag{5.3b}$$

$$\bar{\sigma}_1(t) = \mathbb{E}\left[\sigma(t,\widehat{X}(t))\sigma^{\mathsf{T}}(t,\widehat{X}(t))\right]^{1/2}. \tag{5.3c}$$

On the other hand, the second approach was based on minimising an upper bound to the terminal mean square error

$$\left\|\widetilde{X}_a(T) - \widetilde{X}(T)\right\|^2,$$

which results in the following parameters

$$A(t) = \mathbb{C}\left[a(t,\widehat{X}(t)),\widehat{X}(t)\right]\mathbb{V}\left[\widehat{X}(t)\right]^{-1}, \tag{5.4a}$$

$$b(t) = \mathbb{E}\left[a(t,X(t))\right] - A(t)\mathbb{E}\left[\widehat{X}(t)\right], \tag{5.4b}$$

$$\bar{\sigma}_2(t) = \mathbb{E}\left[\sigma(t,\widehat{X}(t))\right]. \tag{5.4c}$$

These parameters can then be plugged into the filtering and smoothing equations in Section 4.1.2 and if the parameters are obtained by selecting $\widehat{X}(t)$ to be the filtering process then they can be computed online. In this case, the parameters selected by Equation (5.3) gives the standard assumed density filter and smoother as described in Section 4.1.3. However, this does not hold when the parameters are selected by Equation (5.4) unless $\sigma(t,x)=\sigma(t)$ since $\mathbb{E}\big[\sigma(t,\widehat{X}(t))\big] \neq \mathbb{E}\big[\sigma(t,\widehat{X}(t))\sigma^{\mathsf{T}}(t,\widehat{X}(t))\big]^{1/2}$ in general.

In any case, when the linearisation is done with respect to the filtering distribution in the filtering pass, with respect to the smoothing distribution in the smoothing pass, and $\sigma(t,x)=\sigma(t)$ then the parameters due to both Equations (5.3) and (5.4) reduce to the Type I smoother of Särkkä and Sarmavuori (2013) (see Approximation 8). This is Proposition 2 of Tronarp and Särkkä (2019a). On the other hand, if the linearisation with respect to the filtering distribution is kept in the smoothing pass, then the Type II smoother (see Approximation 9) of Särkkä and Sarmavuori (2013) is obtained provided that Equation (5.3) was used for the linearisation.

However, the most important part of the contribution is that the present linearisation scheme can be employed to develop iterative smoothers. This is done by selecting $\big\{\widehat{X}(t)\big\}_{0\leq t\leq T}$ to be the current best approximation to the smoothing process, then linearise using either Equation (5.3) or Equation (5.4), which in turn can be plugged into the filtering and smoothing equations of Section 4.1.2 to obtain a better approximation. Evidence from simulation studies suggests that a moderate to vast improvement in estimation accuracy can be expected and that the iterations converge rapidly.

5.1.2 Reflections and Outlook

It is encouraging that the Type II smoother of Särkkä and Sarmavuori (2013) could be derived with the present linearisation methods. Unfortunately, for the the Type I smoother of Särkkä and Sarmavuori (2013) this is only the case when $\sigma(t,x)=\sigma(t)$. Moreover, if it is possible to derive the Type I smoother with a linearisation approach, then it would imply that the smoothing equations in Approximation 8 can be written as

$$\dot{\mu}^S(t) = \hat{A}(t)\mu^S(t) + \hat{b}(t) + \hat{Q}(t)\big[\Sigma^F(t)\big]^{-1}\big(\mu^S(t)-\mu^F(t)\big), \tag{5.5a}$$

$$\dot{\Sigma}^S(t) = \Big[\hat{A}(t)+\hat{Q}(t)\big[\Sigma^F(t)\big]^{-1}\Big]\Sigma^S(t) + \Sigma^S(t)\Big[\hat{A}(t)+\hat{Q}(t)\big[\Sigma^F(t)\big]^{-1}\Big]^{\mathsf{T}} - \hat{Q}(t). \tag{5.5b}$$

At the present moment it is not clear how this can be done in an appropriate manner.

Another important gap is the convergence analysis of the iterative smoothers, which only consists of fairly limited empirical findings at the present moment. The convergence analysis of the discrete time counterparts has been established by a similar argument as for the Gauss–Newton method (García-Fernández et al., 2017, Tronarp et al., 2018a). However, it is not clear how to extend these arguments to the continuous-discrete-time setting considered here.

5.2 Publication II

In this contribution, the posterior linearisation method is extended to more general state-space models than what was considered by García-Fernández et al. (2015, 2017). More specifically, the following state-space model is considered:

$$X_0 \sim \mathcal{N}\left(\mu_0^F, \Sigma_0^F\right), \tag{5.6a}$$

$$X_n \mid X_{n-1} \sim p_{X_n|X_{n-1}}(x \mid x_{n-1}), \tag{5.6b}$$

$$Y_n \mid X \sim p_{Y_n|X_n}(y \mid x_n). \tag{5.6c}$$

In order to arrive at a tractable state estimator it is required that the following moments are tractable:

$$\begin{aligned}
&\mathbb{E}[X_{n+1} \mid X_n], \\
&\mathbb{V}[X_{n+1} \mid X_n], \\
&\mathbb{E}[Y_n \mid X_n], \\
&\mathbb{V}[Y_n \mid X_n],
\end{aligned}$$

where $\mathbb{E}[X_{n+1} \mid X_n]$ and $\mathbb{V}[X_{n+1} \mid X_n]$ are the expectation and covariance of X_{n+1} with respect to $p_{X_n|X_{n-1}}$, and $\mathbb{E}[Y_n \mid X_n]$ and $\mathbb{V}[Y_n \mid X_n]$ are expectation and covariance of Y_n with respect to $p_{Y_n|X_n}$. Additionally, to arrive at a feasible state estimator Assumption 2 is required.

Assumption 2. *For any μ_n and Σ_n the following holds*

$$\mathbb{C}\left[\mathbb{E}[X_{n+1} \mid X_n], X_n\right] \neq 0,$$
$$\mathbb{C}\left[\mathbb{E}[Y_n \mid X_n], X_n\right] \neq 0,$$

where $\mathbb{C}$ is the cross-covariance with respect to $X_n \sim \mathcal{N}(\mu_n, \Sigma_n)$.

Essentially, Assumption 2 means that a non-trivial linear estimator of X_n using Y_n or X_{n+1} can be constructed. For example, the method does not work on the stochastic volatility model (see, Eq. (3.27)) since

$$\mathbb{C}\left[\mathbb{E}[Y_n \mid X_n], X_n\right] = \mathbb{C}\left[\mathbb{E}[\exp(X_n/2)V_n \mid X_n], X_n\right] = \mathbb{C}[0, X_n] = 0.$$

Consequently, statistical linear regression yields the parameters $\hat{C} = 0$, $\hat{d} = 0$, and $\hat{R} = \mathbb{V}[\exp(X_n)]$, which implies that no update takes place (see Approximations 14 and 17).

5.2.1 The Main Findings

The idea is that the statistical linear regression method can be applied to Equations (5.6), which is most easily realised by defining the following quantities:

$$\begin{aligned}
a_n(x) &= \mathbb{E}[X_n \mid X_{n-1}]\Big|_{X_{n-1}=x},\\
Q_n(x) &= \mathbb{V}[X_n \mid X_{n-1}]\Big|_{X_{n-1}=x},\\
c_n(x) &= \mathbb{E}[Y_n \mid X_n]\Big|_{X_n=x},\\
R_n(x) &= \mathbb{V}[Y_n \mid X_n]\Big|_{X_n=x}.
\end{aligned}$$

That is, applying the statistical linear regression method on the state-space model in Equation (5.6) is equivalent to applying it to a conditionally Gaussian state-space model and an iterative smoother can be implemented by Approximations 16, 17, and 18.

Additionally, the convergence analysis of García-Fernández et al. (2017) was extended to the present setting. While the condition of convergence is hard to verify in practice, the result guarantees that the method is convergent if initialised sufficiently closed to a fixed point. The reader is referred to Tronarp et al. (2018a) for more precise statements regarding the convergence.

5.2.2 Reflections and Outlook

The contribution has already had notable impact on research in Gaussian process classification (García-Fernández et al., 2019a) and target tracking with von Mises–Fisher distributed direction of arrival measurements (García-Fernández et al., 2019b). However, one drawback of the present formulation is that the Kullback–Leibler based Gaussian approximations to $p_{X_{n+1}|X_n}$ and $p_{Y_n|X_n}$ that were used to justify the original method (García-Fernández et al., 2015, 2017) no longer work in the present context.

In order to illustrate this point, let us consider the linearisation of some density $p_{Y|X}(y \mid x)$ with respect to $X \sim p_X(x) = \mathcal{N}(x;\mu,\Sigma)$. The Kullback–Leibler justification of statistical linear regression and indeed posterior linearisation is then to approximate $p_{Y|X}$ by

$$p_{Y|X}(y \mid x) \approx \hat{p}_{Y|X}(y \mid x) = \mathcal{N}\left(y;\hat{C}x+\hat{d},\hat{R}\right), \tag{5.10}$$

the parameters $\hat{C}$, $\hat{d}$, and $\hat{R}$ are then selected to minimise the following Kullback–Leibler divergence

$$(\hat{C},\hat{d},\hat{R}) = \underset{C,d,R}{\arg\min}\, D\left(p_{Y|X}p_X \,||\, \hat{p}_{Y|X}p_X\right). \tag{5.11}$$

However, already at Equation (5.10) problems are afoot. Namely, if Y has outcomes in the space $\mathbb{Y} \neq \mathbb{R}^m$ for all m then $\hat{p}_{Y|X}$ does not define a probability density on $\mathbb{Y}$. Consequently the Kullback–Leibler divergence in Equation (5.11) does not make sense. For example, in the contribution the density $p_{Y|X}(y \mid x) = \mathrm{Po}(y; c\exp(x))$ is considered, hence $\mathbb{Y} = \mathbb{N}_0$.

While the method reduces to iterated extended Kalman filter if $Q_n(x) = Q_n$, $R_n(x) = R_n$ and if the statistical linear regression solutions are approximated with Taylor series Bell (1994), Bell and Cathey (1993) and it can be related to a Kullback–Leibler minimisation for conditionally Gaussian models (García-Fernández et al., 2015, 2017), it is not clear how to appropriately characterise the method in the more general setting. This is an important topic for future research and indeed there are already some results in this direction. Unfortunately, these results are not mature enough for publication at the time of writing this thesis. Lastly, it is noted that the conditional moment tricks used in the present contribution appear similar to the methods of West et al. (1985), however they do not develop iterations. On the other hand, West et al. (1985) deals with general exponential family measurements, which suggests an extension of the present iterative method.

5.3 Publication III

In this contribution, the problem of numerically approximating the solutions of ordinary differential equations was examined. That is,

$$\dot{y}(t) = f(t, y), \quad y(0) = y_0, \tag{5.12}$$

where $f\colon [0,T] \times \mathbb{R}^d$ is the vector field and $y_0 \in \mathbb{R}^d$ is the initial value. The approach was to pose the problem as inference in a probabilistic state-space model, which places it in the field of probabilistic numerics (Hennig et al., 2015). A continuous-time prior $X\colon [0,T] \times \mathbb{R}^{d(q+1)}$ was defined via a stochastic differential equation

$$\mathrm{d}X(t) = FX(t)\,\mathrm{d}t + u\,\mathrm{d}t + D\,\mathrm{d}W(t) \tag{5.13}$$

and each $d \times 1$ sub-vector of X is denoted by $X^{(j)}$, $j = 1, \ldots, q+1$. Here $X^{(1)}$ and $X^{(2)}$ models $y(t)$ and $\dot{y}(t)$, respectively, and consequently their initial conditions are set to $X^{(1)}(0) = y_0$ and $X^{(2)}(0) = f(0, y_0)$. The remaining state components are reasonably used to model higher order derivatives. The q times integrated Wiener process (Eq. (3.13)) is a prominent example of a prior for the solution of an ordinary differential equation (Kersting and Hennig, 2016, Schober et al., 2014, 2019). Furthermore, for any time point t the following likelihood model was used

$$h\big(t, X(t)\big) \triangleq X^{(2)}(t) - f\big(t, X^{(1)}(t)\big), \tag{5.14a}$$

$$Z(t) \mid X \sim \mathcal{N}\big(h(t, X(t)), R\big), \tag{5.14b}$$

$$z(t) \triangleq 0, \tag{5.14c}$$

which essentially means that $X^{(1)}$ is measured to solve the differential equation with some added noise as determined by R.

5.3.1 The Main Findings

For a uniform grid $\{t_n\}_{n=0}^N$ with step size h Equations (5.13) and (5.14) can be written as a discrete-time probabilistic state-space model by using the exact discretisation method (Eq. (3.7)):

$$X(t_{n+1}) \mid X(t_n) \sim \mathcal{N}\big(A(h)X(t_n) + \xi(h), Q(h)\big), \tag{5.15a}$$

$$Z(t_{n+1}) \mid X(t_{n+1}) \sim \mathcal{N}\big(h(t_{n+1}, X(t_{n+1})), R\big), \tag{5.15b}$$

$$z(t_{n+1}) \triangleq 0. \tag{5.15c}$$

Now the problem of numerically solving Equation (5.12) has been put in a form that can be solved by the methods discussed in Section 4.2. That is, assumed density filtering, for which $R = 0$, whereas R is set to some small value for particle filtering, to enable standard linearisation methods for constructing proposal distributions (Doucet et al., 2000) at the cost of only targeting the actual inference problem approximately. If the vector field $f(t, y)$ is affine in y then the measurement function $h(t, x)$ is affine in x, which implies inference can be solved exactly by Kalman filtering and Rauch–Tung–Striebel smoothing (see Section 4.2.2). Furthermore, the stability theory of Kalman filtering in linear time invariant systems (Anderson and Moore, 1979) has a rather remarkable consequence. That is, consider modelling the solution of the following linear test problem

$$\dot{y}(t) = \Lambda y(y), \quad y(0) = y_0 \in \mathbb{R}^d \tag{5.16}$$

with Equation (5.15), where the prior on X is a q times integrated Wiener process, and assume Λ is of full rank. Then the filter mean of $X(t_n)$ tends to zero for any $q = 1, 2, \ldots$, $\mu_n^F \to 0$ as $n \to \infty$ (Tronarp et al., 2019, Theorem 2). That is, the solution estimate of will tend to zero regardless of the stability properties of Equation (5.16). This is a stronger property than A-stability (Dahlquist, 1963), where it is required that the estimate of the solution of Equation (5.16) tends to zero whenever the exact solution does.

5.3.2 Reflections and Outlook

The major point of the contribution is that the solution of Equation (5.12) can be modelled by a probabilistic state-space model and thus approximating the solution of Equation (5.12) is simply a matter of selecting an approximate inference algorithm for Equation (5.15) and some standard options were show cased (Tronarp et al., 2019). However, for each individual algorithm a convergence rate of the solution estimate is required, which is presently lacking. Though due to the connection with estimation of stochastic processes and splines (Kimeldorf and Wahba, 1970, Sidhu and Weinert, 1979, Weinert and Kailath, 1974, Weinert and Sidhu, 1978, Weinert et al., 1979), it would appear that the convergence analysis of splines (Golomb and Jerome, 1971, Wahba, 1973a,b) would carry over to the case of affine vector fields after necessary changes are made.

Lastly, it is rather unfortunate that the present development requires $R > 0$ in Equation (5.15) for the particle filtering approach to work. This is required for the approximations of the locally optimal importance density (see Eq. (4.55)) of Doucet et al. (2000), otherwise the likelihood ratio is zero with probability one

$$\frac{L_n(X_{n,l}) p_{X_n|X_{n-1}}(X_{n,l} \mid X_{n-1,l})}{q_{X_n|X_{0:n-1}}(X_{n,l} \mid X_{n-1,l})} = 0 \quad \text{(with probability one).}$$

This is because $X^{(2)}_{n,l} \neq f(t_n, X_{n,l})$ almost surely under $q_{X_n|X_{0:n-1}}$ as constructed by the methods of Doucet et al. (2000) when $R = 0$. One way to get around this drawback is to construct $q_{X_n|X_{0:n-1}}$ such that $X^{(2)}_{n,l} = f(t_n, X^{(1)}_{n,l})$ with probability one. Though doing this in a clever manner is not entirely trivial.

5.4 Publication IV

The problem examined in this contribution is primarily that of approximate Bayesian updating as discussed in Section 2.1. Suppose that the prior $\pi(x)$ is a member of some parametric set of probability densities, $\mathscr{P} = \{p_\theta : \theta \in \Theta \subseteq \mathbb{R}^m\}$. That is, $\pi = p_{\theta_0}$ for some $p_{\theta_0} \in \mathscr{P}$ and $\theta_0 \in \Theta$. Then for a likelihood $L(x) = m(y \mid x)$, Bayes' rule gives the posterior

$$p(x \mid y) = \frac{L(x)\pi(x)}{\int_{\mathbb{X}} L(x)\pi(x) \mathrm{d}x}.$$

The key idea is to define a curve $\gamma(\tau, x)$, $\tau \in [0,1]$ such that $\gamma(0,x) = \pi(x)$ and $\gamma(1,x) = p(x \mid y)$. Thereafter, the ideas from information geometry and projection filtering (see Sections 2.2 and 4.1.3) are used to approximate γ with some other curve $\hat{\gamma}_{\theta(\tau)}(x)$ such that $\hat{\gamma}_{\theta(\tau)} \in \mathscr{P}$ for all $\tau \in [0,1]$. More specifically, γ was defined according to (Tronarp and Särkkä, 2019b, Eq. (6))

$$\gamma(\tau, x) = \frac{[L(x)]^\tau \pi(x)}{\int_{\mathbb{X}} [L(x)]^\tau \pi(x) \mathrm{d}x}, \tag{5.17}$$

which implies that

$$\partial_\tau \gamma^{1/2}(\tau, x) = \mathscr{U}_{X|Y}\left[\gamma^{1/2}\right](\tau, x), \quad \gamma^{1/2}(0,x) = \pi^{1/2}(x),$$

where the operator $\mathscr{U}_{X|Y} : \mathscr{L}_2 \mapsto \mathscr{L}_2$ is given by[1]

$$\mathscr{U}_{X|Y}[\phi](x) = \frac{1}{2}\left(\ell(x) - \int_{\mathbb{X}} \ell(x)\phi^2(x) \mathrm{d}x\right)\phi(x).$$

---

[1] Strictly speaking, the domain of the operator $\mathscr{U}_{X|Y}$ is not all of $\mathscr{L}_2$ but rather all $p^{1/2} \in \mathscr{L}_2$ for which the integral $\int_{\mathbb{X}} \ell(x)p(x) \mathrm{d}x$ is finite.

By the same method used in projection filtering (see Section 4.1.3) the approximate curve, *the projection update*, is given by

$$\partial_\tau \hat{\gamma}^{1/2}_{\theta(\tau)}(x) = \Pi_{\theta(\tau)} \circ \mathcal{U}_{X|Y}\left[\hat{\gamma}^{1/2}_{\theta(\tau)}\right](x), \quad \gamma^{1/2}(0,x) = \pi^{1/2}(x), \tag{5.18}$$

which in turn defines an approximation of the posterior according to

$$p(x \mid y) \approx \hat{\gamma}_{\theta(1)}(x).$$

As with the projection filter and smoother, Equation (5.18) defines a curve in Θ, which is given by (Tronarp and Särkkä, 2019b, Eq. (8))

$$\dot{\theta}(\tau) = g^{-1}(\theta(\tau))\mathbb{E}_{\theta(\tau)}\left[\ell(X)\nabla_{\theta(\tau)} \log \hat{\gamma}_{\theta(\tau)}(X)\right], \quad \theta(0) = \theta_0, \tag{5.19}$$

where $g(\theta(\tau))$ is the Fisher information matrix of $\mathscr{P}$ evaluated at $\theta(\tau)$ and $\mathbb{E}_{\theta(\tau)}$ is the expectation with respect to $\hat{\gamma}_{\theta(\tau)}$.

5.4.1 The Main Findings

The most important finding is that if $\mathscr{P}$ defines an exponential class that is conjugate to the likelihood $L(x)$. That is $p(x \mid y) \in \mathscr{P}$, then the projection update as defined by Equation (5.18) is exact (Tronarp and Särkkä, 2019b, Theorem 1).

Theorem 4. *Let* $p_\theta(x) = g(x)\exp\left(\theta^\top T(x) - \kappa(\theta)\right)$ *for any* $\theta \in \Theta$ *and* $p_\theta \in \mathscr{P}$ *and assume* $L(x) \propto \exp(\eta^\top T(x))$*. Then*

$$\hat{\gamma}_{\theta(1)}(x) = p(x \mid y) = g(x)\exp\left(\theta_{X|Y}^\top T(x) - \kappa(\theta_{X|Y})\right),$$

where

$$\theta_{X|Y} = \theta + \eta.$$

It is clear that Theorem 4 applies to the Gaussian family of priors as well, since the Gaussian family is also an exponential family (Tronarp and Särkkä, 2019b, Theorem 2).

Another important point is that the expectation in Equation (5.19) can often be computed in closed form. This is exemplified in the article by taking the Gaussian family as prior combined with Laplace likelihoods or the stochastic volatility likelihood. More broadly, for exponential families with sufficient statistic $T(x)$, and log-likelihoods $\ell(x)$ that are polynomials in the sufficient statistic $T(x)$, computing the expectation in Equation (5.19) reduces to computing derivatives of the log-partition function $\kappa(\theta)$.

5.4.2 Reflections and Outlook

The idea of exploiting curves going from priors to posteriors is not new as such. Indeed, it is similar to the concept of tempering, which has been used to develop Monte Carlo samplers based on the sequential Monte

Carlo approach (Del Moral et al., 2006), and the notion of progressive likelihoods that is used to develop progressive Gaussian filters (Steinbring and Hanebeck, 2014).

However, the key idea behind the projection update is not the curve going from prior to posterior per se but rather how it is used to develop approximate inference algorithms. It appears that the formalism of information geometry provides an elegant way of doing this. Furthermore, it may be possible to develop diagnostic tools to assess the accuracy of the approximation by monitoring the local projection residuals (Brigo et al., 1999, Section 6).

An issue with the present development is that the projection update in Equation (5.18) depends on the curve γ, which was selected in a rather ad-hoc manner. There are infinitely many curves from prior to posterior of which γ as defined in Equation (5.17) is just one example. It thus appears fruitful to explore principled ways of selecting the curve connecting the prior to the posterior, perhaps at the cost of losing some of the aforementioned computational benefits of the present selection. One option in this directions would be to consider geodesics on the unit sphere in $\mathscr{L}_2$ (Bauer et al., 2015).

5.5 Publication V

In this work, the problem of tracking a reference vector is examined. That is, there is some time-varying vector $r(t) \in \mathbb{R}^3$ such that $\partial_t \left\| r(t) \right\|^2 = 0$. Without loss of generality, the case when $r(t)$ is on the unit sphere is considered, $r(t) \in \mathbb{S}^2$. The deterministic kinematics are given by

$$\dot{r}(t) = -\Omega(t) \times r(t),$$

where $\Omega(t)$ is the angular velocity. It is assumed that measurements $\breve{\Omega}(t)$ of $\Omega(t)$ are available through a three-axis gyroscope and that $r(t)$ is measured such that the likelihood at every measurement instant t_n is given by

$$L(t_n, r) \propto \exp\Big(-\frac{V(\rho_n^2)}{2}\Big),$$

where

$$\rho_n^2 = \frac{\left\| y(t_n) - gQr(t_n) - b \right\|^2}{\sigma_Y^2},$$

where $g \in \mathbb{R}_+$ is some scalar gain, $Q \in \mathbb{SO}(3)$ is a rotation matrix, $b \in \mathbb{R}^3$ is a bias term, and $V : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is at least once differentiable.

5.5.1 The Main Findings

Firstly, in order to account for the noisy measurements of $\Omega(t)$ a norm-preserving stochastic differential equation for $r(t)$ was formulated, which is driven by $\breve{\Omega}(t)$. It is given by

$$\mathrm{d}R(t) = -\breve{\Omega}(t) \times R(t)\,\mathrm{d}t - \gamma^2 R(t)\,\mathrm{d}t + \gamma R(t) \times \mathrm{d}W(t), \tag{5.20}$$

which can be verified to be norm preserving by using Itô's formula on $\left\|R(t)\right\|^2/2$ (Tronarp et al., 2018b, Lemma 1).

Secondly, an assumed density filter based on the von Mises–Fisher distribution (Mardia and Jupp, 2000) is developed, whose probability density function for $\mathbb{S}^2$ is given by

$$\mathcal{VMF}(r;\mu,\eta) = \exp\left(\eta\mu^\top r - \psi_3(\eta)\right)\chi_{\mathbb{S}^2}(r),$$

where $\eta \in [0,\infty)$, $\mu \in \mathbb{S}^2$, $\chi_{\mathbb{S}^2}$ is an indicator function for the set $\mathbb{S}^2$, and

$$\psi_3(\eta) = \log(4\pi\sinh\eta) - \log\eta.$$

From Equation (5.20) and using the assumed density method, $R(t) \sim \mathcal{VMF}(r;\mu(t),\eta(t))$, the following prediction equations are retrieved

$$\dot{\mu}(t) = -\breve{\Omega}(t) \times \mu(t), \tag{5.22a}$$

$$\dot{\eta}(t) = -\gamma^2 \frac{\psi_3'(\eta(t))}{\psi_3''(\eta(t))}. \tag{5.22b}$$

Furthermore, if $V(\rho_n^2) = \rho_n^2$ (i.e., Gaussian likelihoods) then $\mathcal{VMF}$ is a conjugate prior and the parameter update is given by (Tronarp et al., 2018b, Proposition 1)

$$\eta(t_n) = \left\| \frac{g}{\sigma_Y^2} Q^\top (y-b) + \eta(t_n^-)\mu(t_n^-) \right\|,$$

$$\mu(t_n) = \frac{\frac{g}{\sigma_Y^2} Q^\top (y-b) + \eta(t_n^-)\mu(t_n^-)}{\eta(t_n)}.$$

In general, $\mathcal{VMF}$ is approximately a conjugate prior by Taylor expanding V in ρ_n^2 up to first order around $\rho_n^2\big|_{r(t_n)=\mu(t_n^-)}$ (Tronarp et al., 2018b, Approximation 1).

5.5.2 Reflections and Outlook

Assumed density filters based on the von Mises–Fisher distribution have previously been developed in discrete-time (Bukal et al., 2017, Kurz et al., 2016, Marković et al., 2014a,b, Traa and Smaragdis, 2014). The contribution of the present publication was to develop a continuous-discrete formulation and to provide further evidence that there are benefits to appropriately accounting for the geometry of the state-space $\mathbb{X}$.

As discussed in Section 4.1.3 there is a correspondence between the assumed density filter and the projection filter in the Gaussian setting. There is evidence that this might be the case for the proposed the von Mises–Fisher filter as well. More specifically, the von Mises–Fisher distribution is an exponential family with natural parameter $\theta(t) = \eta(t)\mu(t)$. From Eq. (5.22) together and the chain rule it follows that

$$\dot{\theta}(t) = -\left(\frac{\gamma^2}{\|\theta(t)\|} \frac{\psi'\big(\|\theta(t)\|\big)}{\psi''\big(\|\theta(t)\|\big)} \mathrm{I} + [\breve{\Omega}(t)]_\times \right) \theta(t),$$

which can equivalently be written as

$$\dot{\theta}(t) = \Big[\nabla_{\theta(t)} \nabla_{\theta(t)}^{\mathsf{T}} \psi_3\big(\|\theta(t)\|\big) \Big]^{-1} \mathbb{E}_{\theta(t)}[\mathscr{A} R(t)], \tag{5.24}$$

where $\mathbb{E}_{\theta(t)}$ is the expectation with respect to $R(t) \sim \mathscr{V}\mathscr{M}\mathscr{F}\big(\theta(t)/\|\theta(t)\|, \|\theta(t)\|\big)$ and $\mathscr{A}$ is the generator associated with Equation (5.20), which is to be understood as acting point-wise on $\phi(r) = r$. Equation (5.24) precisely coincides with the projection method of prediction for exponential families, since the sufficient statistic for the von Mises–Fisher distribution is in fact $s(r) = r$ (see e.g., Eq. (15) in Koyama 2018). However, there is a catch, namely that the projection filter operates on the evolution of the probability density as determined by the Fokker-Planck equation and not the evolution of the sufficient statistic, which is how the continuous-discrete von Mises–Fisher filter was derived (Tronarp et al., 2018b). That is, the preceding argument is not sufficient to establish the continuous-discrete von Mises–Fisher filter as a projection filter but it does tempt further investigation.

5.6 Publication VI

In this contribution, the continuous-discrete time inference problem is examined for conditionally Gaussian models with state-independent diffusion matrix

$$\begin{aligned} \mathrm{d}X(t) &= a\big(t, X(t)\big)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W(t), \\ Y(t_n) \mid X &\sim \mathscr{N}\big(c(t_n, x(t_n)), R(t_n)\big). \end{aligned}$$

The problem is approached in the assumed density framework. However, rather than working with the conventional prediction equations (see Eq. (6)), basis expansions of the Wiener process are employed (see Section 3.1.1). This is done interval-wise $[t_{n-1}, t_n], \quad n = 1, \ldots, N$ and $t_0 = 0$ according to

$$\mathrm{d}X(t) = a\big(t, X(t)\big)\,\mathrm{d}t + \sigma(t) \sum_{l=1}^{L} U_{n,l} \phi_{n,l}(t)\,\mathrm{d}t, \quad t \in [t_{n-1}, t_n],$$

where $U_{n,l} \sim \mathscr{N}(0, \mathrm{I})$, $U_{n,l}$ is independent from $U_{n,l'}$ if $l \neq l'$, and $\{\phi_{n,l}\}_{l=1}^{\infty}$ is a basis in $\mathscr{L}_2([t_{n-1}, t_n])$. Consequently, an approximate model on the grid

$\{t_n\}_{n=0}^{N}$ is given by

$$\mathbf{U}_n = \begin{pmatrix} U_{n,1}^{\mathsf{T}} & \dots & U_{n,L}^{\mathsf{T}} \end{pmatrix}^{\mathsf{T}}, \tag{5.26a}$$

$$\breve{a}_n(X(t_{n-1}),\mathbf{U}_n) = X(t_n) + \int_{t_{n-1}}^{t_n} a(t,X(t))\,\mathrm{d}t + \sum_{l=1}^{L}\int_{t_{n-1}}^{t_n} \sigma(t)U_{n,l}\phi_{n,l}(t)\,\mathrm{d}t, \tag{5.26b}$$

$$X(t_n) \approx \breve{a}_n(X(t_{n-1}),\mathbf{U}_n). \tag{5.26c}$$

5.6.1 The Main Findings

Due to the approximation in Equation (5.26c), the filtering problem on the measurement grid $\{t_n\}_{n=1}^{N}$ is of a standard form and can be solved by either numerical integration or Taylor series expansions of $\breve{a}_n$ in both $X(t_{n-1})$ and $\mathbf{U}_n$ (Särkkä, 2013, Algorithm 5.5). The former approach was already presented by Lyons et al. (2014). Furthermore, by either using numerical integration or Taylor series expansions, smoother gains can be computed for the smoothing problem on the measurement grid (Tronarp et al., 2018a, Algorithms 1 and 2). Additionally, a scheme for approximating the smoothing solution between measurement grid points was also developed (Tronarp et al., 2018a, Section III.D). However, it appears to behave fairly poorly (Tronarp et al., 2018a, see, e.g., Fig. 2).

5.6.2 Reflections and Outlook

While it appears that the series expansion approach can serve as a decent alternative to the standard assumed density smoothers on the measurement grid points, it appears to be a poor choice if interpolation of the smoothing solution between measurement grid points is needed. It is presently not clear how to rectify this issue. On the other hand, one may note that the conditional mean and covariance of the dynamics due to the approximation in Equation (5.26c) are given by

$$\begin{aligned}
\mathbb{E}[X(t_n)\,|\,X(t_{n-1})] &= \int_{\mathbb{U}} \breve{a}_n\big(X(t_{n-1},\mathbf{u}_n)\big)\mathcal{N}(\mathbf{u}_n;0,\mathrm{I})\,\mathrm{d}\mathbf{u}_n, \\
\mathbb{V}[X(t_n)\,|\,X(t_{n-1})] &= \int_{\mathbb{U}} \breve{a}_n\big(X(t_{n-1},\mathbf{u}_n)\big)\breve{a}_n^{\mathsf{T}}\big(X(t_{n-1},\mathbf{u}_n)\big)\mathcal{N}(\mathbf{u}_n;0,\mathrm{I})\,\mathrm{d}\mathbf{u}_n \\
&\quad - \Big(\int_{\mathbb{U}} \breve{a}_n\big(X(t_{n-1},\mathbf{u}_n)\big)\mathcal{N}(\mathbf{u}_n;0,\mathrm{I})\,\mathrm{d}\mathbf{u}_n\Big) \\
&\quad \times \Big(\int_{\mathbb{U}} \breve{a}_n\big(X(t_{n-1},\mathbf{u}_n)\big)\mathcal{N}(\mathbf{u}_n;0,\mathrm{I})\,\mathrm{d}\mathbf{u}_n\Big)^{\mathsf{T}}.
\end{aligned}$$

Consequently it is straight-forward to apply the methods developed in Publication II to arrive at an iterative smoother on the measurement grid.

References

2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, May 31 - June 7 2014. IEEE.

J. Ala-Luhtala, S. Särkkä, and R. Piché. Gaussian filtering and variational approximations for Bayesian smoothing in continuous-discrete stochastic dynamic systems. *Signal Processing*, 111:124–136, 2015.

S. Amari. *Differential-Geometrical Methods in Statistics*, volume 28. Springer Science & Business Media, 2012.

S. Amari and H. Nagaoka. *Translations of Mathematical Monographs: Methods of Information Geometry*, volume 191. American Mathematical Society, 2007.

B. D. O. Anderson. Fixed interval smoothing for nonlinear continuous time systems. *Information and Control*, 20(3):294–300, 1972.

B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Information and System Sciences Series. Prentice-Hall, 1979.

I. Arasaratnam and S. Haykin. Cubature Kalman filters. *IEEE Transactions on Automatic Control*, 54(6):1254–1269, 2009.

I. Arasaratnam, S. Haykin, and R. J. Elliott. Discrete-time nonlinear filtering algorithms using Gauss–Hermite quadrature. *Proceedings of the IEEE*, 95(5): 953–977, 2007.

C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. S. Shawe-taylor. Variational inference for diffusion processes. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 17–24. Curran Associates, Inc., Vancouver, B.C., Canada, December 3 - 5 2008.

M. S. Arulampalam, S. Maskell, and N. Gordon. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions Signal Processing*, 50:174–188, 2002.

Y. Bar-Shalom, X. R. Li, and T. Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software*. John Wiley & Sons, 2004.

M. Bauer, S. Joshi, and K. Modin. Diffeomorphic density matching by optimal information transport. *SIAM Journal on Imaging Sciences*, 8(3):1718–1751, 2015.

B. M. Bell. The iterated Kalman smoother as a Gauss–Newton method. *SIAM Journal on Optimization*, 4(3):626–636, 1994.

B. M. Bell and F. W. Cathey. The iterated Kalman filter update as a Gauss–Newton method. *IEEE Transaction on Automatic Control*, 38(2):294–297, 1993.

V. E. Beneš. Exact finite-dimensional filters for certain diffusions with nonlinear drift. *Stochastics: An International Journal of Probability and Stochastic Processes*, 5(1-2):65–92, 1981.

D. Brigo, B. Hanzon, and F. LeGland. A differential geometric approach to nonlinear filtering: the projection filter. *IEEE Transactions on Automatic Control*, 43(2):247–252, 1998.

D. Brigo, B. Hanzon, and F. Le Gland. Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli*, 5(3):495–534, 1999.

M. Bukal, I. Marković, and I. Petrović. Score matching based assumed density filtering with the von Mises–Fisher distribution. In *20th International Conference on Information Fusion* ISI (2017).

O. Cappé, E. Moulines, and T. Rydé. *Inference in Hidden Markov Models*. Springer, 2005.

O. Cappé, S. J. Godsill, and E. Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(2): 899–924, 2007.

J. L. Crassidis and J. L. Junkins. *Optimal Estimation of Dynamic Systems*. Chapman & Hall/CRC, 2004.

D. Crisan and A. Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3): 736–746, 2002.

B. Cseke, D. Schnoerr, M. Opper, and G. Sanguinetti. Expectation propagation for continuous time stochastic processes. *Journal of Physics A: Mathematical and Theoretical*, 49(49):494002, 2016.

G. Dahlquist. A special stability problem for linear multistep methods. *BIT Numerical Mathematics*, 3(1):27–43, 1963.

F. E. Daum. Exact finite dimensional nonlinear filters for continuous time processes with discrete time measurements. In *The 23rd IEEE Conference on Decision and Control*, pages 16–22, Las Vegas, Nevada, USA, December 12 - 14 1984. IEEE.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

R. Douc and O. Cappé. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pages 64–69, Zagreb, Croatia, September 15 - 17 2005. IEEE.

A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.

A. Doucet, N. De Freitas, and N. Gordon. An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

Á. F. García-Fernández, L. Svensson, and M. R. Morelande. Iterated statistical linear regression for Bayesian updates. In *17th International Conference on Information Fusion* ISI (2014), pages 1–8.

Á. F. García-Fernández, L. Svensson, M. R. Morelande, and S. Särkkä. Posterior linearization filter: Principles and implementation using sigma points. *IEEE Transactions on Signal Processing*, 63(20):5561–5573, 2015.

Á. F. García-Fernández, L. Svensson, and S. Särkkä. Iterated posterior linearisation smoother. *IEEE Transactions on Automatic Control*, 62(4):2056–2063, 2017.

Á. F. García-Fernández, F. Tronarp, and S. Särkkä. Gaussian process classification using posterior linearization. *IEEE Signal Processing Letters*, 26(5):735–739, 2019a.

Á. F. García-Fernández, F. Tronarp, and S. Särkkä. Gaussian target tracking with direction-of-arrival von Mises–Fisher measurements. *IEEE Transactions on Signal Processing*, 67(11):2960–2972, 2019b.

A. Gelb. *Applied Optimal Estimation*. MIT press, 1974.

I. Gilitschenski, G. Kurz, , S. J. Julier, and U. D. Hanebeck. Efficient Bingham filtering based on saddlepoint approximations. In *17th International Conference on Information Fusion* ISI (2014).

I. Gilitschenski, G. Kurz, S. J. Julier, and U. D. Hanebeck. Unscented orientation estimation based on the Bingham distribution. *IEEE Transactions on Automatic Control*, 61(1):172–177, 2016.

J. Glover and L. P. Kaelbling. Tracking the spin on a ping pong ball with the quaternion Bingham filter. In *2014 IEEE International Conference on Robotics and Automation (ICRA)* ICR (2014), pages 4133–4140.

S. J. Godsill and P. J. Rayner. *Digital Audio Restoration: A Statistical Model Based Approach*. Springer, 1998.

S. J. Godsill, A. Doucet, and M. West. Monte Carlo smoothing for nonlinear time series. *Journal of the American Statistical Association*, 99(465):156–168, 2004.

M. Golomb and J. Jerome. Linear ordinary differential equations with boundary conditions on arbitrary point sets. *Transactions of the American Mathematical Society*, 153:235–264, 1971.

G. H. Golub and J. H. Welsch. Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23(106):221–230, 1969.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing*, 140:107–113, 1993.

F. Gustafsson and G. Hendeby. Some relations between extended and unscented Kalman filters. *IEEE Transactions on Signal Processing*, 60(2):545–555, 2012.

P. Hennig, M. A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142, 2015.

Y. Huang, Y. Zhang, N. Li, S. M. Naqvi, and J. Chambers. A robust Student's t based cubature filter. In *19th International Conference on Information Fusion* ISI (2016), pages 9–16.

17th International Conference on Information Fusion, Salamanca, Spain, July 7 - 10 2014. ISIF, IEEE.

19th International Conference on Information Fusion, Heidelberg, Germany, July 5 - 8 2016. ISIF, IEEE.

20th International Conference on Information Fusion, Xi'an, China, July 10 - 13 2017. ISIF, IEEE.

K. Ito and K. Xiong. Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control*, 45(5):910–927, May 2000.

A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.

S. J. Julier and J. K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.

S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte. A new approach for filtering nonlinear systems. In *Proceedings of the 1995 American Control Conference*, pages 1628–1632, Seattle, WA, USA, June 21 - 23 1995. AACC, IEEE.

S. J. Julier, J. K. Uhlmann, and H. 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):477–482, Mar 2000.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 1988.

T. Karvonen and S. Särkkä. Classical quadrature rules via Gaussian processes. In *27th International Workshop on Machine Learning for Signal Processing*, Tokyo, Japan, September 25 - 28 2017. IEEE, IEEE.

H. Kersting and P. Hennig. Active uncertainty calibration in Bayesian ODE solvers. In *Uncertainty in Artificial Intelligence (UAI) 2016*, New York City, NY, USA, June 25 - 29 2016. AUAI.

G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*, volume 23. Springer Science & Business Media, 2013.

S. Koyama. Projection smoothing for continuous and continuous-discrete stochastic dynamic systems. *Signal Processing*, 144:333–340, 2018.

G. Kurz, I. Gilitschenski, S. Julier, and U. D. Hanebeck. Recursive estimation of orientation based on the Bingham distribution. In *16th International Conference on Information Fusion*, Istanbul, Turkey, July 9 - 12 2013. ISIF, IEEE.

G. Kurz, I. Gilitschenski, S. Julier, and U. D. Hanebeck. Recursive Bingham filter for directional estimation involving 180 degree symmetry. *Journal on Advances in Information Fusion*, 9(2):90–105, 2014.

G. Kurz, I. Gilitschenski, and U. D. Hanebeck. Unscented von Mises–Fisher filtering. *IEEE Signal Processing Letters*, 23(4):463–467, 2016.

T. Lee. Bayesian attitude estimation with approximate matrix Fisher distributions on SO(3). In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5319–5325, Miami Beach, FL, USA, December 17 - 19 2018a. IEEE.

T. Lee. Bayesian attitude estimation with the matrix Fisher distribution on SO(3). *IEEE Transactions on Automatic Control*, 63(10):3377–3392, 2018b.

T. Lefebvre, H. Bruyninckx, and J. De Schuller. Comment on “A new method for the nonlinear transformation of means and covariances in filters and estimators” [with authors’ reply]. *IEEE Transactions on Automatic Control*, 47(8):1406–1409, 2002.

. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

C. T. Leondes, J. B. Peller, and E. B. Stear. Nonlinear smoothing theory. *IEEE Transactions on System Science and Cybernetics*, 6(1):63–71, 1970.

X. R. Li and V. P. Jilkov. Survey of maneuvering target tracking. Part I. dynamic models. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1333–1364, 2003.

E. Lindström, H. Madsen, and J. N. Nielsen. *Statistics for Finance*. Chapman and Hall/CRC, 2015.

W. Luo. *Wiener chaos expansion and numerical solutions of stochastic partial differential equations*. PhD thesis, California Institute of Technology, 2006.

S. M. J. Lyons, S. Särkkä, and A. J. Storkey. Series expansion approximations of Brownian motion for non-linear Kalman filtering of diffusion processes. *IEEE Transactions on Signal Processing*, 62(6):1514–1524, 2014.

K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, 2000.

I. Marković, M. Bukal, J. Ćesić, and I. Petrović. Direction-only tracking of moving objects on the unit sphere via probabilistic data association. In *17th International Conference on Information Fusion* ISI (2014).

I. Marković, F. Chaumette, and I. Petrović. Moving object detection, tracking and following using an omnidirectional camera on a mobile robot. In *2014 IEEE International Conference on Robotics and Automation (ICRA)* ICR (2014), pages 5630–5635.

P. S. Maybeck. *Stochastic Models, Estimation and Control*, volume 1-3. Academic Press, 1979,1982,1982.

J. McNamee and F. Stenger. Construction of fully symmetric numerical integration formulas. *Numerische Mathematik*, 10:327–344, 1967.

B. Øksendal. *Stochastic Differential Equations - An Introduction with Applications*. Springer, 2003.

G. C. Price and D. Williams. Rolling with “slipping”: I. *Séminaire de probabilités de Strasbourg*, 17:194–197, 1983.

J. Prüher and O. Straka. Gaussian process quadrature moment transform. *IEEE Transactions on Automatic Control*, 63(9):2844–2854, 2017.

J. Prüher, F. Tronarp, T. Karvonen, S. Särkkä, and O. Straka. Student-t process quadratures for filtering of non-linear systems with heavy-tailed noise. In *20th International Conference on Information Fusion* ISI (2017).

J. Prüher and M. Šimandl. Bayesian quadrature variance in sigma-point filtering. In J. Filipe, K. Madani, O. Gusikhin, and J. Sasiadek, editors, *Informatics in Control, Automation and Robotics 12th International Conference, ICINCO, Revised Selected Papers*, pages 355–370, Colmar, France, July 21 - 23 2015. Springer.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine learning*. MIT Press, 2006.

H. E. Rauch, F. Tung, and C. T. Striebel. Maximum likelihood estimates of linear dynamic system. *AIAA Journal*, 3(8):1445–1450, Aug 1965.

L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes and Martingales*, volume 1,2,3. Cambridge university press, 2000.

M. Roth, E. Özkan, and F. Gustafsson. A Student's t filter for heavy tailed process and measurement noise. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5770–5774, Vancouver, B.C, Canada, May 26 - 31 2013. IEEE.

S. Särkkä. Unscented Rauch–Tung–Striebel Smoother. *IEEE Transactions on Automatic Control*, 53(3):845–849, April 2008.

S. Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

S. Särkkä and J. Hartikainen. On Gaussian optimal smoothing of non-linear state space models. *IEEE Transactions on Automatic Control*, 55(8):1938–1941, 2010.

S. Särkkä and J. Sarmavuori. Gaussian filtering and smoothing for continuous-discrete dynamic systems. *Signal Processing*, 93:500–510, 2013.

S. Särkkä and A. Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

S. Särkkä and T. Sottinen. Application of Girsanov theorem to particle filtering of discretely observed continuous-time non-linear systems. *Bayesian Analysis*, 3 (3):555–584, 2008.

S. Särkkä, J. Hartikainen, L. Svensson, and F. Sandblom. On the relation between Gaussian process quadratures and sigma-point methods. *arXiv preprint arXiv:1504.05994*, 2015.

M. Schober, D. K. Duvenaud, and P. Hennig. Probabilistic ODE solvers with Runge-Kutta means. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 739–747, Montréal, Canada, December 8 - 13 2014. Curran Associates, Inc.

M. Schober, S. Särkkä, and P. Hennig. A probabilistic model for the numerical solution of initial value problems. *Statistics and Computing*, 29(1):99–122, January 2019.

G. S. Sidhu and H. L. Weinert. Vector-valued Lg-splines I. interpolating splines. *Journal of Mathematical Analysis and Applications*, 70(2):505–529, 1979.

D. Simon. *Optimal State Estimation: Kalman, H_∞, and Nonlinear Approaches*. John Wiley & Sons, 2006.

M. Skoglund, G. Hendeby, and D. Axehill. Extended Kalman filter modifications based on an optimization view point. In *18th International Conference on Information Fusion*, Washington, DC, USA, July 6 - 9 2015. ISIF, IEEE.

J. Steinbring and U. D. Hanebeck. Progressive Gaussian filtering using explicit likelihoods. In *17th International Conference on Information Fusion* ISI (2014).

L. D. Stone, R. L. Streit, T. L. Corwin, and K. L. Bell. *Bayesian Multiple Target Tracking*. Artech House, 2014.

C. T. Striebel. Partial differential equations for the conditional distribution of a Markov process given noisy observations. *Journal of Mathematical Analysis and Applications*, 11:151–159, 1965.

T. Sutter, A. Ganguly, and H. Koeppl. A variational approach to path estimation and parameter inference of hidden diffusion processes. *The Journal of Machine Learning Research*, 17(1):6544–6580, 2016.

D. H. Titterton and J. L. Weston. *Strapdown Inertial Navigation Technology*. The Institute of Electrical Engineers, 2004.

J. Traa and P. Smaragdis. Multiple speaker tracking with the factorial von Mises–Fisher filter. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Reims, France, September 21 - 24 2014. IEEE.

F. Tronarp and S. Särkkä. Iterative statistical linear regression for Gaussian smoothing in continuous-time non-linear stochastic dynamic systems. *Signal Processing*, 159:1–12, 2019a.

F. Tronarp and S. Särkkä. Updates in Bayesian filtering by continuous projections on a manifold of densities. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5032–5036, Brighton, United Kingdom, May 12 - 17 2019b. IEEE.

F. Tronarp, R. Hostettler, and S. Särkkä. Sigma-point filtering for nonlinear systems with non-additive heavy-tailed noise. In *19th International Conference on Information Fusion* ISI (2016).

F. Tronarp, A. F. Garcia-Fernandez, and S. Särkkä. Iterative filtering and smoothing in non-linear and non-Gaussian systems using conditional moments. *IEEE Signal Processing Letters*, 25(3):408–412, March 2018a. ISSN 1070-9908. doi: 10.1109/LSP.2018.2794767.

F. Tronarp, R. Hostettler, and S. Särkkä. Continuous-discrete von Mises–Fisher filtering on S^2 for reference vector tracking. In *21st International Conference on Information Fusion*, pages 1345–1352, Cambridge, United Kingdom, July 10 - 13 2018b. ISIF, IEEE.

F. Tronarp, H. Kersting, S. Särkkä, and P. Hennig. Probabilistic solutions to ordinary differential equations as non-linear Bayesian filtering: A new perspective. *Statistics and Computing, to appear.*, 2019.

M. Van Den Berg and J. T. Lewis. Brownian motion on a hypersurface. *Bulletin of the London Mathematical Society*, 17(2):144–150, 1985.

G. Wahba. A class of approximate solutions to linear operator equations. *Journal of Approximation Theory*, 9(1):61–77, 1973a.

G. Wahba. Convergence rates of certain approximate solutions to Fredholm integral equations of the first kind. *Journal of Approximation Theory*, 7(2): 167–185, 1973b.

H. L. Weinert and T. Kailath. Stochastic interpretations and recursive algorithms for spline functions. *The Annals of Statistics*, 2(4):787–794, 1974.

H. L. Weinert and G. S. Sidhu. A stochastic framework for recursive computation of spline functions–part I: Interpolating splines. *IEEE Transactions on Information Theory*, 24(1):45–50, 1978.

H. L. Weinert, U. B. Desai, and G. S. Sidhu. Arma Splines, System Inverses, and Least-Squares Estimates. *SIAM Journal on Control and Optimization*, 17(4): 525–536, 1979.

M. West, P. J. Harrison, and H. S. Migon. Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, 80(389): 73–83, 1985.

E. Wong and M. Zakai. On the convergence of ordinary integrals to stochastic integrals. *The Annals of Mathematical Statistics*, 36(5):1560–1564, 1965.

Y. Wu, D. Hu, M. Wu, and X. Hu. A numerical-integration perspective on Gaussian filters. *IEEE Transactions on Signal Processing*, 54(8):2910–2921, Aug 2006.

Errata

Publication III

j3 The legend entry ”KF“ in Fig. 5 should be ”EKF“.

Publication IV

The collection of tangent vectors at θ is given by $\{\partial_{\theta_i} p_\theta^{1/2}\}_{i=1}^{m}$ and not $\{\partial_{\theta_i} p_\theta^{1/2}\}_{i=1}^{p}$. "For the projection update (PU) we take $R_n V_n$ [...]“ should be "For the projection update (PU) we take V_n [...]“.

Publication VI

c3 In the unnumbered equation succeeding Eq. (3) it should be $\mathbb{C}^S[\cdot,\cdot] = \mathbb{C}[\cdot,\cdot \mid \mathscr{Y}(t_K)]$ and not $\mathbb{C}^S[\cdot] = \mathbb{E}[\cdot,\cdot \mid \mathscr{Y}(t_K)]$.

Aalto-DD 12/2020

ISBN 978-952-60-8921-8 (printed)
ISBN 978-952-60-8922-5 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Electrical Engineering and Automation
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS