Linköping studies in science and technology. Dissertations.
No. 1530

Particle filters and Markov chains for learning of dynamical systems

Fredrik Lindsten

Department of Electrical Engineering
Linköping University, SE–581 83 Linköping, Sweden

Linköping 2013

Cover illustration: Sample path from a Gibbs sampler targeting a three-dimensional standard normal distribution.

Linköping studies in science and technology. Dissertations.
No. 1530

Particle filters and Markov chains for learning of dynamical systems

Fredrik Lindsten

`lindsten@isy.liu.se`
`www.control.isy.liu.se`
Division of Automatic Control
Department of Electrical Engineering
Linköping University
SE–581 83 Linköping
Sweden

ISBN 978-91-7519-559-9 ISSN 0345-7524

Copyright © 2013 Fredrik Lindsten

Printed by LiU-Tryck, Linköping, Sweden 2013

To Åsa

Abstract

Sequential Monte Carlo (SMC) and Markov chain Monte Carlo (MCMC) methods provide computational tools for systematic inference and learning in complex dynamical systems, such as nonlinear and non-Gaussian state-space models. This thesis builds upon several methodological advances within these classes of Monte Carlo methods.

Particular emphasis is placed on the combination of SMC and MCMC in so called particle MCMC algorithms. These algorithms rely on SMC for generating samples from the often highly autocorrelated state-trajectory. A specific particle MCMC algorithm, referred to as particle Gibbs with ancestor sampling (PGAS), is suggested. By making use of backward sampling ideas, albeit implemented in a forward-only fashion, PGAS enjoys good mixing even when using seemingly few particles in the underlying SMC sampler. This results in a computationally competitive particle MCMC algorithm. As illustrated in this thesis, PGAS is a useful tool for both Bayesian and frequentistic parameter inference as well as for state smoothing. The PGAS sampler is successfully applied to the classical problem of Wiener system identification, and it is also used for inference in the challenging class of non-Markovian latent variable models.

Many nonlinear models encountered in practice contain some tractable substructure. As a second problem considered in this thesis, we develop Monte Carlo methods capable of exploiting such substructures to obtain more accurate estimators than what is provided otherwise. For the filtering problem, this can be done by using the well known Rao-Blackwellized particle filter (RBPF). The RBPF is analysed in terms of asymptotic variance, resulting in an expression for the performance gain offered by Rao-Blackwellization. Furthermore, a Rao-Blackwellized particle smoother is derived, capable of addressing the smoothing problem in so called mixed linear/nonlinear state-space models. The idea of Rao-Blackwellization is also used to develop an online algorithm for Bayesian parameter inference in nonlinear state-space models with affine parameter dependencies.

Populärvetenskaplig sammanfattning

Matematiska modeller av dynamiska förlopp används inom i stort sett alla tekniska och naturvetenskapliga discipliner. Till exempel, inom epidemiologi används modeller för att prediktera, dvs. förutsäga, spridningen av influensavirus inom en population. Antag att vi gör regelbundna observationer av hur många personer i populationen som är smittade. Baserat på denna information kan en modell användas för att prediktera antalet nya sjukdomsfall under, låt säga, nästkommande veckor. Den här typen av information möjliggör att en epidemi kan identifieras i ett tidigt skede, varpå åtgärder kan tas för att minska dess påverkan. Ett annat exempel är att prediktera hur hastigheten och orienteringen på ett flygplan påverkas då en viss styrsignal ställs ut på rodren, vilket är viktigt vid styrsystemdesign. Sådana prediktioner kräver en modell av flygplanets dynamik. Ytterligare ett exempel är att prediktera utvecklingen på en aktiekurs baserat på tidigare observationer. Bör vi helt enkelt anta att kursen imorgon är densamma som idag, eller bör vi även beakta tidigare observationer för att ta hänsyn till eventuella trender? Den typen av frågor besvaras av en modell. Modellen beskriver hur vi ska väga samman den tillgängliga informationen för att kunna göra så bra prediktioner som möjligt.

Användandet av dynamiska modeller spelar således en viktig roll. Det är därför även viktigt att ha tillgång till verktyg för att bygga dessa modeller. Den här avhandlingen behandlar problemet att utnyttja insamlad data för att finna statistiska modeller som beskriver dynamiska förlopp. Detta problem kallas för *systemidentifiering* eller för *statistisk inlärning*. Baserat på exemplen ovan är det lätt att inse att dynamiska modeller används inom vitt skilda områden. Trots detta så är den bakomliggande matematiken i mångt och mycket densamma. Av den anledningen så behandlas inte något specifikt användningsområde i denna avhandling. Istället fokuserar vi på matematiken – de metoder som presenteras kan sedan användas inom ett brett spektra av tillämpningar.

I många fall är det otillräckligt att använda enkla modeller som endast baseras på, till exempel, linjära trender. Inom ekonomi är det vanligt att volatiliteten, dvs. graden av variation, hos en finansiell tillgång varierar med tiden. För att beskriva detta krävs en statistisk modell som kan förändras över tiden. Inom epidemiologi är det viktigt att ha modeller som kan ta hänsyn till det tydliga säsongsberoendet hos ett influensaförlopp. Detta kräver att modellerna innehåller olinjära funktioner som kan beskriva sådana variationer. För att kunna modellera denna typ av *komplexa dynamiska fenomen* så krävs, i någon mening, komplexa matematiska modeller. Detta leder dock till att det statistiska inlärningsproblemet blir matematiskt invecklat – i praktiken till den grad att det inte går att lösa exakt. Detta kan hanteras på två olika sätt. Antingen gör man avkall på flexibiliteten och noggrannheten i modellen, eller så väljer man att ta fram en approximativ lösning till inlärningsproblemet.

I den här avhandlingen följer vi det sistnämnda alternativet. Mer specifikt så används en klass av approximativa inlärningsmetoder som kallas för *Monte Carlo-metoder*. Namnet är en anspelning på det kända kasinot i Monte Carlo och syftar på att dessa metoder baseras på slumptal. För att illustrera konceptet, antag att du lägger en patiens, dvs. ett enmanskortspel som syftar till att lägga ut korten enligt vissa spelregler vilket leder antingen till vinst eller till förlust. Att på förhand räkna ut vad sannolikheten för vinst är kräver komplicerade kombinatoriska uträkningar, som lätt blir praktiskt taget omöjliga att utföra. Ett mer pragmatiskt sätt är att spela ut korten, säg, 100 gånger och notera hur många av dessa försök som resulterar i vinst. Vinstfrekvensen blir en naturlig skattning av vinstsannolikheten. Denna skattning är inte helt tillförlitlig eftersom den är baserad på slumpen, men ju fler försök som utförs, desto högre noggrannhet uppnås i skattningen.

Detta är ett exempel på en enkel Monte Carlo-metod. De metoder som används för att skatta dynamiska modeller är mer invecklade, men grundprincipen är densamma. Metoderna är datorprogram som, genom att generera ett stort antal slumptal, kan skatta intressanta kvantiteter som är omöjliga att beräkna exakt. Detta kan till exempel vara värden på modellparametrar eller sannolikheten att en parameter ligger inom ett visst intervall.

I den här avhandlingen används i huvudsak två klasser av Monte Carlo-metoder, partikelfilter och Markovkedjor. Partikelfiltret är ett systematiskt sätt att utvärdera och uppdatera ett antal slumpmässigt genererade hypoteser. Låt oss återigen betrakta en epidemiologisk modell för influensaprediktion. I praktiken finns ingen exakt vetskap om hur stor del av populationen som är smittad vid ett visst tillfälle. De observationer som görs av antalet insjuknade är av olika anledningar osäkra. Ett partikelfilter kan användas för att hantera denna osäkerhet och skatta det underliggande *tillståndet*, dvs. det faktiskta antalet smittade personer. Detta görs genom att slumpvis generera en mängd hypoteser om hur många personer som är insjuknade. Dessa hypoteser kallas för *partiklar*, därav namnet på metoden. Baserat på de faktiska observationer som görs kan sannolikheterna för de olika hypoteserna utvärderas. De hypoteser som ej är troliga kan avfärdas, medan de mer sannolika hypoteserna dupliceras. Eftersom influensan är ett dynamiskt förlopp, dvs. den förändras över tiden, så måste även hypoteserna uppdateras. Detta görs genom att utnyttja en modell över influensaförloppets dynamik. Dessa två steg upprepas sekventiellt över tiden och partikelfiltret kallas därför för en sekventiell Monte Carlo-metod.

Markovkedjor ligger till grund för en annan klass av Monte Carlo-metoder. En Markovkedja är en sekvens av slumptal där varje tal i sekvensen är statistiskt beroende av det föregående talet. Inom Monte Carlo används Markovkedjor för att generera en sekvens av hypoteser rörande, till exempel, värden på okända modellparametrar. Varje hypotes baseras på den föregående. Systematiska tekniker används för att uppdatera hypoteserna så att de efter hand resulterar i en korrekt modell.

Bidraget i den här avhandlingen är utvecklingen av nya metoder, baserade på partikelfilter och Markovkedjor, som kan användas för att lösa det statistiska inlärningsproblemet i komplexa dynamiska modeller. Partikelfilter och Markovkedjor kan även kombineras, vilket resulterar i än mer kraftfulla metoder som har kommit att kallas för PMCMC-metoder (Particle Markov Chain Monte Carlo). Dessa ligger till grund för en stor del av avhandlingen. I synnerhet presenteras en ny typ av PMCMC-metod som har visat sig vara effektiv jämfört med tidigare alternativ. Som nämnts ovan kan metoden användas inom vitt skilda vetenskapliga områden. Flera variationer och utökningar av den föreslagna metoden presenteras också. Vi tittar även närmre på en specifik klass av dynamiska modeller som kallas för *betingat linjära*. Dessa modeller innehåller en viss struktur, och vi undersöker hur denna struktur kan utnyttjas för att underlätta det statistiska inlärningsproblemet.

Acknowledgments

When I look back at the years that I have spent in the group of Automatic Control (henceforth abbreviated RT), the first thought that comes to mind is *why was there a Peruvian giraffe on a backstreet in Brussels?* Unfortunately, I don't think that we will ever know the truth. The second thing that comes to my mind, however, is that *those years at RT have been lots of fun.* They have been fun for many reasons – the great atmosphere in the group, the fact that I have had so many good friends as colleagues, the pub nights, the barbecues.... However, I also want to acknowledge all the members of the group who, skillfully and with great enthusiasm, engage in every activity whether it's research, teaching or something else. To me, this has been very motivating. Let me spend a few lines on expressing my gratitude to some of the people who have made the years at RT, and the years before, a really great time.

First of all, I would like to thank my supervisor Prof Thomas B. Schön for all the guidance and support. Your dedication and enthusiasm is admirable and I have very much enjoyed working with you. Having you as a supervisor has resulted in (what I think is) a very nice research collaboration and a solid friendship, both which I know will continue in the future. I would also like to thank my co-supervisors Prof Lennart Ljung and Prof Fredrik Gustafsson for providing me with additional guidance and valuable expertise. Thanks to Prof Svante Gunnarsson and to Ninna Stensgård for creating such a nice atmosphere in the group and for always being very helpful.

I am most grateful for the financial support from the projects Calibrating Nonlinear Dynamical Models (Contract number: 621-2010-5876) funded by the Swedish Research Council and CADICS, a Linnaeus Center also funded by the Swedish Research Council.

In 2012, I had the privilege of spending some time as a visiting student researcher at the University of California, Berkeley. I want to express my sincere gratitude to Prof Michael I. Jordan for inviting me to his lab. The four months that I spent in Berkeley were interesting, fun and rewarding on so many levels. I made many new friends and started up several collaborations which I value highly today. Furthermore, it was during this visit that I got the opportunity to work on the theory which is now the foundation for this thesis!

I have also had the privilege of working with several other colleagues from outside the walls of RT. I want to express my thanks to Prof Eric Moulines, Dr Alexandre Bouchard-Côté, Dr Bonnie Kirkpatrick, Johan Wågberg, Roger Frigola, Dr Carl E. Rasmussen, Pete Bunch, Prof Simon J. Godsill, Ehsan Taghavi, Dr Lennart Svensson, Dr Adrian Wills, Prof Brett Ninness and Dr Jimmy Olsson for the time you have spent on our joint papers and/or research projects.

Among my colleagues at RT I want to particularly mention Johan Dahlin, Christian Andersson Naesseth and Dr Emre Özkan, with whom I have been working in close collaboration lately. Johan has also proofread the first part of this thesis. Thank you! Thanks also to my former roommate Lic (soon-to-be-Dr) Zoran Sjanic for being hilarious and for joining me in the creation of the most awesome music quiz even known to man! Thanks to Lic (even-sooner-to-be-Dr) Daniel Petersson for always taking the time to discuss various problems that I have encountered in my research, despite the fact that it has been quite different from yours. Lic Martin Skoglund, Lic André Carvalho Bittencourt and Lic Peter Rosander have made sure that long-distance running is not as lonely as in Alan Sillitoe's story. Thanks! Thanks also to Lic Sina Khoshfetrat Pakazad for being such a good friend and for keeping all the uninvited guests away from *Il Kraken*. Thanks to Floyd – by the way, where are you?

I also want to thank my old (well, we are starting to get old) friends from Y04C, my brother Mikael Lindsten and my other friends and family. My parents Anitha Lindsten and Anders Johansson have my deepest gratitude. I am quite sure that neither *particle filters* nor *Markov chains* mean a thing to you. Despite this you have always been very supportive and I love you both!

Most of all, I want to thank my wonderful wife Åsa Lindsten. Thank you for always believing in me and for all your encouragement and support. To quote Peter Cetera,

> *You're the inspiration!*

I love you!

Finally, in case I have forgotten someone, I would like to thank ________________ for
(your name here)
____________________________. Thanks!
(fill in the reason, e.g., "all the good times")

Linköping, September 2013
Fredrik Lindsten

Contents

Notation **xv**

I Background

1 Introduction **3**
- 1.1 Models of dynamical systems . . . 3
- 1.2 Inference and learning . . . 5
- 1.3 Contributions . . . 7
- 1.4 Publications . . . 8
- 1.5 Thesis outline . . . 10
  - 1.5.1 Outline of Part I . . . 10
  - 1.5.2 Outline of Part II . . . 10

2 Learning of dynamical systems **13**
- 2.1 Modeling . . . 13
- 2.2 Maximum likelihood . . . 15
- 2.3 Bayesian learning . . . 16
- 2.4 Data augmentation . . . 18
- 2.5 Online learning . . . 20

3 Monte Carlo methods **23**
- 3.1 The Monte Carlo idea . . . 23
- 3.2 Rejection Sampling . . . 25
- 3.3 Importance sampling . . . 27
- 3.4 Particle filters and Markov chains . . . 29
- 3.5 Rao-Blackwellization . . . 31

4 Concluding remarks **33**
- 4.1 Conclusions and future work . . . 33
- 4.2 Further reading . . . 34

Bibliography **37**

II Publications

A Backward simulation methods for Monte Carlo statistical inference 45
- 1 Introduction . . . 47
  - 1.1 Background and motivation . . . 48
  - 1.2 Notation and definitions . . . 49
  - 1.3 A preview example . . . 49
  - 1.4 State-space models . . . 52
  - 1.5 Parameter learning in SSMs . . . 53
  - 1.6 Smoothing recursions . . . 54
  - 1.7 Backward simulation in linear Gaussian SSMs . . . 56
  - 1.8 Outline . . . 59
- 2 Monte Carlo preliminaries . . . 59
  - 2.1 Sequential Monte Carlo . . . 59
  - 2.2 Markov chain Monte Carlo . . . 64
- 3 Backward simulation for state-space models . . . 70
  - 3.1 Forward filter/backward simulator . . . 70
  - 3.2 Analysis and convergence . . . 76
  - 3.3 Backward simulation with rejection sampling . . . 80
  - 3.4 Backward simulation with MCMC moves . . . 85
  - 3.5 Backward simulation for maximum likelihood inference . . . 89
- 4 Backward simulation for general sequential models . . . 91
  - 4.1 Motivating examples . . . 91
  - 4.2 SMC revisited . . . 94
  - 4.3 A general backward simulator . . . 96
  - 4.4 Rao-Blackwellized FFBSi . . . 101
  - 4.5 Non-Markovian latent variable models . . . 105
  - 4.6 From state-space models to non-Markovian models . . . 105
- 5 Backward simulation in particle MCMC . . . 109
  - 5.1 Introduction to PMCMC . . . 109
  - 5.2 Particle Marginal Metropolis-Hastings . . . 111
  - 5.3 PMMH with backward simulation . . . 117
  - 5.4 Particle Gibbs with backward simulation . . . 120
  - 5.5 Particle Gibbs with ancestor sampling . . . 128
  - 5.6 PMCMC for maximum likelihood inference . . . 134
  - 5.7 PMCMC for state smoothing . . . 137
- 6 Discussion . . . 137
- Bibliography . . . 140

B Ancestor Sampling for Particle Gibbs 151
- 1 Introduction . . . 153
- 2 Sequential Monte Carlo . . . 154
- 3 Particle Gibbs with ancestor sampling . . . 155
- 4 Truncation for non-Markovian state-space models . . . 158
- 5 Application areas . . . 159
  - 5.1 Rao-Blackwellized particle smoothing . . . 159

5.2 Particle smoothing for degenerate state-space models 160
5.3 Additional problem classes 161
6 Numerical evaluation 161
6.1 RBPS: Linear Gaussian state-space model 161
6.2 Random linear Gaussian systems with rank deficient process noise covariances 162
7 Discussion 163
A Proof of Proposition 1 163
Bibliography 165

C Bayesian semiparametric Wiener system identification 167
1 Introduction 169
2 A Bayesian semiparametric model 171
2.1 Alt. I – Conjugate priors 171
2.2 Alt. II – Sparsity-promoting prior 172
2.3 Gaussian process prior 173
3 Inference via particle Gibbs sampling 174
3.1 Ideal Gibbs sampling 174
3.2 Particle Gibbs sampling 175
4 Posterior parameter distributions 178
4.1 MNIW prior – Posterior of Γ and Q 178
4.2 GH prior – Posterior of Γ, Q and $\bar{\tau}$ 178
4.3 Posterior of r 180
4.4 Posterior of $h(\,\cdot\,)$ 180
4.5 Posterior of η 181
5 Convergence analysis 181
5.1 Convergence of the Markov chain 182
5.2 Consistency of the Bayes estimator 184
6 Numerical illustrations 185
6.1 6th-order system with saturation 186
6.2 4th-order system with non-monotone nonlinearity 187
6.3 Discussion 188
7 Conclusions and future work 189
A Choosing the hyperparameters 190
Bibliography 191

D An efficient SAEM algorithm using conditional particle filters 195
1 Introduction 197
2 The EM, MCEM and SAEM algorithms 198
3 Conditional particle filter SAEM 199
3.1 Markovian stochastic approximation 200
3.2 Conditional particle filter with ancestor sampling 200
3.3 Final identification algorithm 203
4 Numerical illustration 203
5 Conclusions 205
Bibliography 207

E Rao-Blackwellized particle smoothers for mixed linear/nonlinear SSMs 209
1 Introduction 211
2 Background 212
    2.1 Particle filtering and smoothing 212
    2.2 Rao-Blackwellized particle filtering 213
3 Rao-Blackwellized particle smoothing 213
4 Proofs 216
5 Numerical results 217
6 Conclusion 219
Bibliography 221

F A non-degenerate RBPF for estimating static parameters in dynamical models 223
1 Introduction 225
2 Degeneracy of the RBPF – the motivation for a new approach 227
3 A non-degenerate RBPF for models with static parameters 229
    3.1 Sampling from the marginals 230
    3.2 Gaussian mixture approximation 230
    3.3 Resulting Algorithm 232
4 Numerical illustration 232
5 Discussion and future work 233
6 Conclusions 234
Bibliography 237

G An explicit variance reduction expression for the Rao-Blackwellised particle filter 239
1 Introduction and related work 241
2 Background 243
    2.1 Notation 243
    2.2 Particle filtering 243
    2.3 Rao-Blackwellised particle filter 244
3 Problem formulation 245
4 The main result 247
5 Relationship between the proposals kernels 248
    5.1 Example: Bootstrap kernels 249
6 Discussion 250
7 Conclusions 251
A Proofs 251
Bibliography 253

Notation

PROBABILITY

Notation	Meaning
$\sim$	Sampled from or distributed according to
$\mathbb{P}, \mathbb{E}$	Probability, expectation
$\mathrm{Var}, \mathrm{Cov}$	Variance, covariance
$\xrightarrow{\mathrm{D}}$	Convergence in distribution
$\mathcal{L}(X \in \cdot\,)$	Law of the random variable X
$\|\mu_1 - \mu_2\|_{\mathrm{TV}}$	Total variation distance, $\sup_A \lvert\mu_1(A) - \mu_2(A)\rvert$

COMMON DISTRIBUTIONS

Notation	Meaning
$\mathrm{Cat}(\{p_i\}_{i=1}^n)$	Categorical over $\{1, \ldots, n\}$ with probabilities $\{p_i\}_{i=1}^n$
$\mathcal{U}([a, b])$	Uniform over the interval $[a, b]$
$\mathcal{N}(m, \Sigma)$	Multivariate Gaussian with mean m and covariance Σ
δ_x	Point-mass at x (Dirac δ-distribution)

OPERATORS AND OTHER SYMBOLS

Notation	Meaning
$\cup, \cap$	Set union, intersection
$\mathrm{card}(S)$	Cardinality of the set S
S^{c}	Complement of S in Ω (given by the context)
$I_S(\,\cdot\,)$	Indicator function of set S
I_d	d-dimensional identity matrix
A^{T}	Transpose of matrix A
$\det(A)$, $\lvert A\rvert$	Determinant of matrix A
$\mathrm{tr}(A)$	Trace of matrix A

$\text{vec}(A)$	Vectorization, stacks the columns of A into a vector
$\text{diag}(v)$	Diagonal matrix with elements of v on the diagonal
$\otimes$	Kronecker product
$\text{supp}(f)$	Support of function f, $\{x : f(x) > 0\}$
$\|f\|_\infty$	Supremum norm, $\sup_x \|f(x)\|$
$\text{osc}(f)$	Oscillator norm, $\sup_{(x,x')} \|f(x) - f(x')\|$
$a_{m:n}$	Sequence, $(a_m, a_{m+1}, \ldots, a_n)$
$\triangleq$	Definition

ABBREVIATIONS

Abbreviation	**Meaning**
ACF	Autocorrelation function
ADM	Average derivative method
APF	Auxiliary particle filter
ARD	Automatic relevance determination
a.s.	almost surely
CLGSS	Conditionally linear Gaussian state-space
CLT	Central limit theorem
CPF	Conditional particle filter
CPF-AS	Conditional particle filter with ancestor sampling
CSMC	Conditional sequential Monte Carlo
DPMM	Dirichlet process mixture model
ESS	Effective sample size
EM	Expectation maximization
FFBSi	Forward filter/backward simulator
FFBSm	Forward filter/backward smoother
FIR	Finite impulse response
GH	Generalized hyperbolic
GIG	Generalized inverse-Gaussian
GMM	Gaussian mixture model
GP	Gaussian process
GPB	Generalized pseudo-Bayesian
HMM	Hidden Markov model
i.i.d.	independent and identically distributed
IMM	Interacting multiple model
IW	Inverse Wishart
JMLS	Jump Markov linear system
JSD	Joint smoothing density
KF	Kalman filter
KLD	Kullback-Leibler divergence
LGSS	Linear Gaussian state-space
LTI	Linear time-invariant
MBF	Modified Bryson-Frazier
MCEM	Monte Carlo expectation maximization

MCMC	Markov chain Monte Carlo
MH	Metropolis-Hastings
MH-FFBP	Metropolis-Hastings forward filter/backward proposing
MH-FFBSi	Metropolis-Hastings forward filter/backward simulator
MH-IPS	Metropolis-Hastings improved particle smoother
ML	Maximum likelihood
MLE	Maximum likelihood estimator
MNIW	Matrix normal inverse Wishart
MPF	Marginal particle filter
MRF	Markov random field
PDF	Probability density function
PEM	Prediction-error method
PF	Particle filter
PG	Particle Gibbs
PGAS	Particle Gibbs with ancestor sampling
PGBS	Particle Gibbs with backward simulation
PIMH	Particle independent Metropolis-Hastings
PMCMC	Particle Markov chain Monte Carlo
PMMH	Particle marginal Metropolis-Hastings
PSAEM	Particle stochastic approximation expectation maximization
PSEM	Particle smoother expectation maximization
RB-FFBSi	Rao-Blackwellized forward filter/backward simulator
RB-FF/JBS	Rao-Blackwellized forward filter/joint backward simulator
RB-F/S	Rao-Blackwellized filter/smoother
RBMPF	Rao-Blackwellized marginal particle filter
RBPF	Rao-Blackwellized particle filter
RBPS	Rao-Blackwellised particle smoother
RMSE	Root-mean-square error
RS	Rejection sampling
RS-FFBSi	Rejection sampling forward filter/backward simulator
RTS	Rauch-Tung-Striebel
SAEM	Stochastic approximation expectation maximization
SIR	Susceptible/infected/recovered
SMC	Sequential Monte Carlo
SSM	State-space model
TV	Total variation

Part I

Background

1

Introduction

This thesis addresses inference and learning of dynamical systems. Problems lacking closed form solutions are considered and we therefore make use of computational statistical methods based on random simulation to address these problems. In this introductory chapter, we formulate and motivate the learning problem which is studied throughout the thesis.

1.1 Models of dynamical systems

An often encountered problem in a wide range of scientific fields is to make predictions about some dynamical process based on previous observations from the process. As an example, in the field of epidemiology the evolution of contagious diseases is studied (Keeling and Rohani, 2007). Seasonal influenza epidemics each year cause millions of severe illnesses and hundreds of thousands of deaths worldwide (Ginsberg et al., 2009). Furthermore, new strains of influenza viruses can result in pandemic situations with very severe effects on the public health. In order to minimize the harm caused by an epidemic or a pandemic situation, a problem of paramount importance is to be able to predict the spread of the disease. Assume that regular observations are made of the number of infected individuals within a population, e.g. through disease case reports. Alternatively, Ginsberg et al. (2009) have demonstrated that this type of information can be acquired by monitoring search engine query data. Using these observations, we wish to predict how many new cases of illness that will occur within the population, say, during the coming weeks. The ability to accurately make such predictions can enable early response to epidemic situations, which in turn can reduce their impact.

There are numerous other areas in which similar prediction problems for dynamical processes arise. In finance, the ability to predict the future price of an asset based on previous recordings of its value is of key relevance (Hull, 2011) and in automatic control, predictions of how a controlled plant responds to specific commands are needed for efficient control systems design (Åström and Murray, 2008; Ljung, 1999). Additional examples include automotive safety systems (Eskandarian, 2012), population dynamics (Turchin, 2003) and econometrics (Greene, 2008), to mention a few.

Despite the apparent differences between these examples, they can all be studied within a common mathematical framework. We collectively refer to these processes as *dynamical systems*. The word *dynamical* refers to the fact that these processes are evolving over time. For a thorough elementary introduction to dynamical systems, see e.g. the classical text books by Oppenheim et al. (1996) and Kailath (1980).

Common to the dynamical systems studied in this thesis is that observations, or measurements, y_t can be recorded at consecutive time points indexed by $t = 1, 2, \dots$. Based on these readings, we wish to draw conclusions about the system which generated the measurements. For instance, assuming that we have recorded the values $y_{1:t} \triangleq (y_1, \dots, y_t)$, the one-step prediction problem amounts to estimating what the value of y_{t+1} will turn out to be. Should we simply assume that y_{t+1} will be close to the most recent recording y_t, or should we make use of older measurements as well, to account for possible trends? Such questions can be answered by using a *model* of the dynamical system. The model describes how to weigh the available information together to make as good predictions as possible.

For most applications, it is not possible to find models that exactly describe the measurements. There will always be fluctuations and variations in the data, not accounted for by the model. To incorporate such random components, the measurement sequence can be viewed as a realisation of a discrete-time stochastic process. A model of the system is then the same thing as a model of the stochastic process.

A specific class of models, known as state-space models (SSMs), is commonly used in the context of dynamical systems. These models play a central role in this thesis. The structure of an SSM can be seen as influenced by the notion of a physical system. The idea is that, at each time point, the system is assumed to be in a certain *state*. The state contains all relevant information about the system, i.e. if we would know the state of the system we would have full insight into its internal condition. However, the state is typically not known. Instead, we measure some quantities which depend on the state in some way. To exemplify the idea, let x_t be a random variable representing the state at time t. An SSM for the system could then, for instance, be given by,

$$x_{t+1} = a(x_t) + v_t, \tag{1.1a}$$

$$y_t = c(x_t) + e_t. \tag{1.1b}$$

The expression (1.1a) describes the evolution of the system state over time. The state at time $t+1$ is given by a transformation of the current state $a(x_t)$, plus some *process noise* v_t. The process noise accounts for variations in the system state, not accounted for by the model. Equation (1.1a) describes the dynamical evolution of the system and it is therefore known as the dynamic equation. The second part of the model, given by (1.1b), describes how the measurement y_t depends on the state x_t and some *measurement noise* e_t. Consequently, (1.1b) is called the measurement equation. The model of a dynamical system specified by (1.1) thus consists of the functions a and c, but also of the noise characteristics, i.e. of the probability distributions for the process noise and the measurement noise. The concept of SSMs will be further discussed in Chapter 2 and in Section 1 of Paper A.

1.2 Inference and learning

As argued above, models of dynamical systems are of key relevance in many scientific disciplines. Hence, it is crucial to have access to tools with which these models can be built. In this thesis, we consider the problem of *learning* models of dynamical systems based on available observations. On a high level, the learning problem can be described as follows,

> **Learning:** Based on observations of the process $\{y_t\}_{t\geq 1}$, find a mathematical model which, without being too complex, as accurately as possible can describe the observations.

A complicating factor when addressing this problem is that the state process $\{x_t\}_{t\geq 1}$ in (1.1) is unobserved; it is said to be latent or hidden. Instead, as recognized in the description above, any conclusions that we wish to draw regarding the system must be inferred from observations of the measurement sequence $\{y_t\}_{t\geq 1}$. A task which is tightly coupled to the learning problem is therefore to draw inference about the latent state,

> **State inference:** Given a fully specified SSM and based on observations $\{y_t\}_{t\geq 1}$, draw conclusions about some past, present or future state of the system, which is not directly visible but related to the measurements through the model.

For instance, even if the system model would be completely known, making a prediction about a future value of the system state amounts to solving a state inference problem. As we shall see in Chapter 2, state inference often plays an important role as an intermediate step when addressing the learning problem.

There exists a wide variety of models and modeling techniques. One common approach is to make use of parametric models. That is, the SSM in (1.1) is specified only up to some unknown (possibly multi-dimensional) parameter, denoted θ. The learning problem then amounts to draw inference about the value of θ based on data collected from the system. This problem is studied in several related scientific fields, e.g. statistics, system identification and machine learning, all with their own notation and nomenclature. We will mainly use the word *learning*, but we also refer to this problem as *identification*, *parameter inference* and *parameter estimation*. We provide an example of a parametric SSM below. Alternative modeling techniques are discussed in more detail in Chapter 2. See also the monographs by Cappé et al. (2005), Ljung (1999) and West and Harrison (1997) for a general treatment of the learning problem in the context of dynamical systems.

Example 1.1

To describe the evolution of a contagious disease, a basic epidemiological model is the susceptible/infected/recovered (SIR) model (Keeling and Rohani, 2007). In a population of constant size, we let S_t, I_t and R_t represent the fractions of susceptible, infected and recovered individuals at time t, respectively. Rasmussen et al. (2011) and Lindsten and

Schön (2012) study a time-discrete SIR model with environmental noise and seasonal fluctuations, which is given by

$$S_{t+1} = S_t + \mu - \mu S_t - \beta_t S_t I_t v_t, \tag{1.2a}$$

$$I_{t+1} = I_t - (\gamma + \mu) I_t + \beta_t S_t I_t v_t, \tag{1.2b}$$

$$R_{t+1} = R_t + \gamma I_t - \mu R_t. \tag{1.2c}$$

Here, β_t is a seasonally varying transmission rate given by $\beta_t = \bar{\beta}(1 + \alpha \sin(2\pi t/365))$, where it is assumed that the time t is measured in days. Together with α and $\bar{\beta}$, the parameters of the model are the birth/death rate μ, the recovery rate γ and the variance σ_v^2 of the zero-mean Gaussian process noise v_t. That is, we can collect the system parameters in a vector

$$\theta = \begin{pmatrix} \alpha & \bar{\beta} & \mu & \gamma & \sigma_v^2 \end{pmatrix}^{\mathsf{T}}.$$

The SIR model in (1.2) corresponds to the process model (1.1a). Note that the system state $x_t = (S_t, I_t, R_t)$ is not directly observed. Instead, Lindsten and Schön (2012) consider an observation model which is inspired by the Google Flu Trends project (Ginsberg et al., 2009). The idea is to use the frequency of influenza related search engine queries to infer knowledge about the dynamics of the epidemic. The observation model, corresponding to (1.1b), is a linear relationship between the observations and the log-odds of infected individuals, i.e.

$$y_t = \log\left(\frac{I_t}{1 - I_t}\right) + e_t, \tag{1.3}$$

with e_t being a zero-mean Gaussian noise.

Lindsten and Schön (2012) use a method denoted as particle Gibbs with backward simulation (PGBS; see Section 5 of Paper A) to learn the parameters of this SIR model. Using the identified model, a state inference problem is solved in order to make one-month-ahead predictions of the number of infected individuals. The results from a simulation study are shown in Figure 1.1, illustrating the possibility of forecasting the disease activity by using a dynamical model.

The SIR model (1.2) is an example of an SSM in which the functions a and c in (1.1) depend nonlinearly on the state x_t. Such SSMs are referred to as *nonlinear*. Conversely, if both a and c are linear (of affine) functions of x_t, the SSM is also called linear. Linear models play an important role for many applications. However, there are also many cases in which they are inadequate for capturing the dynamics of the system under study; the epidemiological model above being one example. Despite this limitation, much emphasis has traditionally been put on linear models. One factor contributing to this is that nonlinear models by nature are much more difficult to work with. However, as we develop more sophisticated computational tools and acquire more and more computational resources, we can also address increasingly more challenging problems. Inspired by this fact, this thesis is focused on the use of computational methods for inference and learning of nonlinear dynamical models.

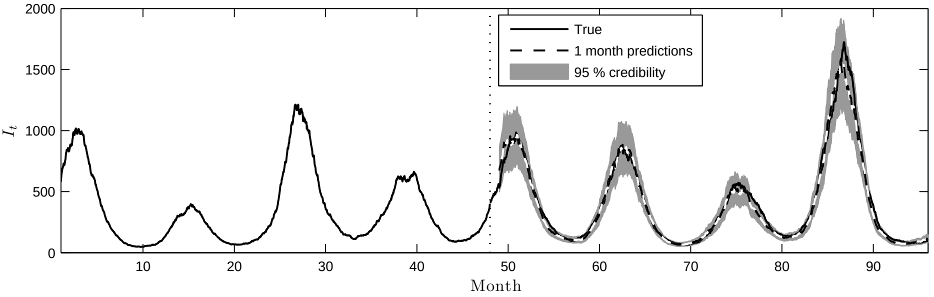

Figure 1.1: *Number of infected individuals I_t in a population of size 10^6 over an 8 year period. Data from the first 4 years are used to learn the unknown parameters of the model. For the consecutive 4 years, one-month-ahead predictions are computed using the estimated model. See (Lindsten and Schön, 2012) for details on the experiment.*

In particular, we make use of a class of methods based on random simulation, referred to as Monte Carlo methods (Robert and Casella, 2004; Liu, 2001). This is a broad class of computational algorithms which are useful for addressing high-dimensional, intractable integration problems. We make use of Monte Carlo methods to address both state inference and learning problems. In particular, we employ methods based on so called Markov chains and on interacting particle systems. An introduction to basic Monte Carlo methods is given in Chapter 3. More advanced methods are discussed in Paper A in Part II of this thesis.

1.3 Contributions

The main contribution of this thesis is the development of new methodology for state inference and learning of dynamical systems. In particular, an algorithm referred to as particle Gibbs with ancestor sampling (PGAS) is proposed. It is illustrated that PGAS is a useful tool for both Bayesian and frequentistic learning as well as for state inference. The following contributions are made:

- The PGAS sampler is derived and its validity assessed by viewing the individual steps of the algorithm as a sequence of partially collapsed Gibbs steps (Paper B).
- A truncation strategy for backward sampling in so called non-Markovian latent variable models is developed and used together with PGAS (Paper B). The connections between non-Markovian models and several important types of SSMs are discussed (Paper A), motivating the development of inference strategies for this model class.
- An algorithm based on PGAS is developed for the classical problem of Wiener system identification (Paper C).
- PGAS is combined with stochastic approximation expectation maximization, resulting in method for frequentistic learning of nonlinear SSMs (Paper D).

Many nonlinear models encountered in practice contain some tractable substructure. When addressing learning and inference problems for such models, this structure can be exploited to improve upon the performance of the algorithms. In this thesis we consider a type of structure exploitation referred to as *Rao-Blackwellization.* We develop and analyse several Rao-Blackwellized Monte Carlo methods for inference and learning in nonlinear SSMs. The following contributions are made:

- A Rao-Blackwellized particle smoother is developed for a class of mixed linear/nonlinear SSMs (Paper E).
- An online, Bayesian identification algorithm, based on the Rao-Blackwellized particle filter, is developed (Paper F).
- The asymptotic variance of the Rao-Blackwellized particle filter is analysed and an expression for the variance reduction offered by Rao-Blackwellization is derived (Paper G).

1.4 Publications

Published work of relevance to this thesis are listed below in reversed chronological order. Items marked with a star are included in Part II of the thesis.

⋆ F. Lindsten and T. B. Schön. Backward simulation methods for Monte Carlo statistical inference. *Foundations and Trends in Machine Learning*, 6(1):1–143, 2013.

⋆ F. Lindsten, T. B. Schön, and M. I. Jordan. Bayesian semiparametric Wiener system identification. *Automatica*, 49(7):2053–2063, 2013b.

⋆ F. Lindsten. An efficient stochastic approximation EM algorithm using conditional particle filters. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

⋆ F. Lindsten, P. Bunch, S. J. Godsill, and T. B. Schön. Rao-Blackwellized particle smoothers for mixed linear/nonlinear state-space models. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013a.

J. Dahlin, F. Lindsten, and T. B. Schön. Particle Metropolis Hastings using Langevin dynamics. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

E. Taghavi, F. Lindsten, L. Svensson, and T. B. Schön. Adaptive stopping for fast particle smoothing. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

⋆ F. Lindsten, M. I. Jordan, and T. B. Schön. Ancestor sampling for particle Gibbs. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS) 25*, pages 2600–2608. 2012a.

⋆ F. Lindsten, T. B. Schön, and L. Svensson. A non-degenerate Rao-Blackwellised particle filter for estimating static parameters in dynamical models. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, July 2012c.

F. Lindsten, T. B. Schön, and M. I. Jordan. A semiparametric Bayesian approach to Wiener system identification. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, July 2012b.

J. Dahlin, F. Lindsten, T. B. Schön, and A. Wills. Hierarchical Bayesian ARX models for robust inference. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, July 2012.

A. Wills, T. B. Schön, F. Lindsten, and B. Ninness. Estimation of linear systems using a Gibbs sampler. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, July 2012.

F. Lindsten and T. B. Schön. On the use of backward simulation in the particle Gibbs sampler. In *Proceedings of the 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012.

⋆ F. Lindsten, T. B. Schön, and J. Olsson. An explicit variance reduction expression for the Rao-Blackwellised particle filter. In *Proceedings of the 18th IFAC World Congress*, Milan, Italy, August 2011b.

F. Lindsten and T. B. Schön. Identification of mixed linear/nonlinear state-space models. In *Proceedings of the 49th IEEE Conference on Decision and Control (CDC)*, Atlanta, USA, December 2010.

Other publications, loosely connected to the material presented in this thesis, are:

F. Lindsten, H. Ohlsson, and L. Ljung. Clustering using sum-of-norms regularization; with application to particle filter output computation. In *Proceedings of the IEEE Workshop on Statistical Signal Processing (SSP)*, Nice, France, June 2011a.

F. Lindsten, J. Callmer, H. Ohlsson, D. Törnqvist, T. B. Schön, and F. Gustafsson. Geo-referencing for UAV navigation using environmental classification. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Anchorage, USA, May 2010.

F. Lindsten, P.-J. Nordlund, and F. Gustafsson. Conflict detection metrics for aircraft sense and avoid systems. In *Proceedings of the 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes (SafeProcess)*, Barcelona, Spain, July 2009.

1.5 Thesis outline

The thesis is divided into two parts. The first part contains background material and an introduction to the problem studied throughout the thesis. The second part is a compilation of seven edited publications. However, the first publication, Paper A, is a self-contained tutorial article covering many of the topics studied in the thesis. Paper A should therefore be viewed as part of the introduction, complementing the material presented in Part I.

1.5.1 Outline of Part I

Chapter 2 introduces the learning problem for dynamical systems. The maximum likelihood and the Bayesian learning criteria are defined and we discuss the basic strategies for addressing these problems. Chapter 3 is an introduction to basic Monte Carlo methods. The algorithms discussed in this chapter are the building blocks needed for constructing more advanced methods later in the thesis. Readers familiar with Monte Carlo statistical inference can skip this chapter. Finally, Chapter 4 concludes the thesis and point out possible directions for future work.

1.5.2 Outline of Part II

Part II is a compilation of seven edited publications.

Paper A,

> F. Lindsten and T. B. Schön. Backward simulation methods for Monte Carlo statistical inference. *Foundations and Trends in Machine Learning*, 6(1):1–143, 2013.

is a self-contained tutorial article covering a branch of Monte Carlo methods referred to as *backward simulators*. These methods are useful for inference in probabilistic models containing latent stochastic processes, e.g. SSMs. The first two sections of this paper should preferably be read as part of the introduction, as they complement the background material presented in Part I of the thesis. In particular,

1. SSMs are introduced in Chapter 2, but a more thorough discussion is provided in Section 1 of Paper A.
2. Particle filters and Markov chains, the two main computational tools which are employed throughout this thesis, are briefly discussed in Chapter 3. However, a more thorough introduction is given in Section 2 of Paper A.

In the remaining sections of Paper A, several Monte Carlo methods based on particle filters and on Markov chains are discussed. In particular, it is illustrated how backward simulation can be used to address the so called smoothing problem and many state-of-the-art *particle smoothers* are surveyed.

Paper B,

> F. Lindsten, M. I. Jordan, and T. B. Schön. Ancestor sampling for particle Gibbs. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and

> K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS) 25*, pages 2600–2608. 2012a.

contains the derivation of the PGAS method. PGAS belongs to the family of so called particle Markov chain Monte Carlo (PMCMC) algorithms. PMCMC is a combination of particle filters and Markov chain theory, resulting in potent tools for Bayesian learning and state inference. PGAS makes use of a technique reminiscent of backward simulation, albeit implemented in a forward-only fashion, to improve the performance of the algorithm. In particular, PGAS has been found to work well even when using few particles in the underlying particle filter. This implies that the algorithm is computationally competitive when compared with many other particle-filter-based methods. It is also discussed how PGAS can be used for inference in the challenging class of non-Markovian latent variable models.

Paper C,

> F. Lindsten, T. B. Schön, and M. I. Jordan. Bayesian semiparametric Wiener system identification. *Automatica*, 49(7):2053–2063, 2013b.

makes use of PGAS for addressing the classical problem of Wiener system identification. A Wiener system is composed of a linear dynamical system followed by a static nonlinearity. That is, the measured quantity is a nonlinear transformation of the output from the linear dynamical system. A semiparametric model is assumed for the Wiener system. The model consists of a parametric model for the linear dynamical system and a nonparametric model for the static nonlinearity. The resulting identification algorithm can handle challenging situations, such as process noise and non-monotonicity of the nonlinearity, in a systematic manner.

Paper D,

> F. Lindsten. An efficient stochastic approximation EM algorithm using conditional particle filters. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

is also based on the PGAS algorithm. In its original formulation, PGAS is useful for addressing the Bayesian learning problem. In this paper, the algorithm is adapted to instead solve the maximum likelihood problem. This is accomplished by using PGAS together with, so called, stochastic approximation expectation maximization. The resulting algorithm is shown to be computationally very competitive when compared with alternative particle-filter-based expectation maximization methods.

The last three papers are not (directly) related to PGAS. Instead, the common denominator in these papers is that they make use of Rao-Blackwellization. Many nonlinear models encountered in practice contain some tractable substructure. In the context of particle filtering, Rao-Blackwellization refers to the process of exploiting such substructures to improve the performance of the algorithms.

Paper E,

> F. Lindsten, P. Bunch, S. J. Godsill, and T. B. Schön. Rao-Blackwellized particle smoothers for mixed linear/nonlinear state-space models. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013a.

presents a Rao-Blackwellized backward simulation method. This algorithm can be used to address the state inference problem in a class of SSMs referred to as mixed linear/nonlinear. In these models, the state can be partitioned into two components, one which enters linearly and one which enters nonlinearly. By exploiting this structure, the proposed algorithm results in more accurate estimators than what is obtained otherwise.

Paper F,

> F. Lindsten, T. B. Schön, and L. Svensson. A non-degenerate Rao-Blackwellised particle filter for estimating static parameters in dynamical models. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, July 2012c.

considers the problem of online Bayesian learning. That is, we seek to learn a model which is continuously updated as new information is collected from the system. Inspired by the Rao-Blackwellized particle filter (RBPF), an approximate method capable of addressing this challenging problem is proposed. The method is applicable for Gaussian models with a linear dependence on the model parameters, but a possibly nonlinear dependence on the system state. At each time point, the posterior distribution of the system parameters is approximated by a Gaussian mixture. The components of this mixture distribution are systematically updated as new information becomes available by using moment matching.

Paper G,

> F. Lindsten, T. B. Schön, and J. Olsson. An explicit variance reduction expression for the Rao-Blackwellised particle filter. In *Proceedings of the 18th IFAC World Congress*, Milan, Italy, August 2011b.

the final paper of the thesis, provides an analysis of the RBPF. By considering the asymptotic variances of the particle filter and the RBPF, respectively, an expression for the improvement offered by Rao-Blackwellization is obtained.

2

Learning of dynamical systems

This chapter introduces the learning problem for dynamical systems. We define the maximum likelihood and the Bayesian learning criteria and discuss the technique of *data augmentation*.

2.1 Modeling

On a high level, we can distinguish between different strategies for building models of dynamical systems as being *white-*, *gray-* or *black-box* modeling techniques (Ljung, 1999). A white-box model is based solely on first principles, such as Newton's laws of motion. A gray-box model is constructed using similar insight into the structure of the dynamical system, but it also contains unknown parameters. These parameters have to be estimated from observations taken from the system. Finally, a black-box model is constructed using only observed data, with no structural knowledge about the system. Black-box models thus have to be flexible in order to capture different types of dynamical phenomena which are present in the data.

For gray- and black-box models, the process of estimating unknown model quantities based on observed data is what we refer to as *learning*. It should be noted, however, that *learning* sometimes refers to a more general problem, including how to specify the model structure, how to the design experiments for data collection etc. However, we shall restrict our attention to the aforementioned subtask, i.e. to estimate parameters or other unknown model quantities once the model structure has been specified.

As pointed out in Chapter 1, we will primarily be concerned with SSMs. This is a comprehensive and flexible class of models of dynamical systems. The additive noise model (1.1) is an example of an SSM. More generally, we can express the model in terms of probability density functions (PDFs) as,

$$x_{t+1} \sim f_\theta(x_{t+1} \mid x_t), \tag{2.1a}$$

$$y_t \sim g_\theta(y_t \mid x_t), \tag{2.1b}$$

with the initial state x_1 distributed according to $\mu_\theta(x_1)$. Here, $f_\theta(x_{t+1} \mid x_t)$ is a Markov kernel encoding the probability of moving from state x_t at time t to state x_{t+1} at time $t+1$. Similarly, $g_\theta(y_t \mid x_t)$ denotes the probability density of obtaining an observation y_t, given that the current state of the system is x_t. The latent state process $\{x_t\}_{t\geq 1}$ is Markov and, conditionally on x_t, the observation y_t is independent of past and future states and observations. SSMs are further discussed and exemplified in Section 1 of Paper A.

Remark 2.1. In the system identification literature (see e.g. Ljung (1999)), particular emphasis is put on learning of dynamical systems used in control applications. Hence, it is common to let the system be excited by some known control input $\{u_t\}_{t\geq 1}$, i.e. by adding a dependence on u_t on the right hand side of (2.1). In this thesis, we will not make such dependence explicit, but this is purely for notational convenience. The learning methods that we consider are indeed applicable also in the presence of a known input signal.

The model (2.1) is said to be *parametric*, since it is specified only up to some finite-dimensional parameter $\theta \in \Theta$, where Θ denotes a set of plausible parameters. As noted in Chapter 1, an often encountered problem is to make predictions about some future output from the system. Based on the model (2.1), the PDF of the one-step predictor can be computed as,

$$p_\theta(y_{t+1} \mid y_{1:t}) = \int g_\theta(y_{t+1} \mid x_{t+1}) p_\theta(x_{t+1} \mid y_{1:t})\, dx_{t+1}, \tag{2.2}$$

where $y_{1:t} = (y_1, \ldots, y_t)$ denotes the observations collected up to time t. There are two things that are interesting to note about this expression. First, the predictor depends on the model parameter θ. Hence, to be able to use the model for making predictions, we need to obtain knowledge about its unknown parameters. Second, the expression (2.2) depends on the predictive density for the latent state $p_\theta(x_{t+1} \mid y_{1:t})$. Consequently, making a prediction about a future output from the system amounts to solving a state inference problem.

The complexity and flexibility of a parametric model is typically related to the dimensionality of θ, i.e. to the number of adjustable parameters. However, there is a trade-off between using many parameters to obtain an expressive model, and using few parameters to unambiguously being able to learn the values of these parameters. If the model is too simplistic to capture the dynamics of the system, we say that it suffers from *under-fitting*. On the contrary, if the model is too complex and thereby prevents accurate learning of the model parameters, there is a problem of *over-fitting*. Over- and under-fitting occurs when there is a mismatch between the model complexity and the amount of available data, or more precisely the amount of information available in the data.

A different take on modeling of dynamical systems is provided by *nonparametric* models. The word *nonparametric* does not imply that these models lack parameters. On the contrary, it means that the number of parameters is allowed to grow with the amount of data. Mathematically, this is accomplished by allowing for an infinite-dimensional latent structure in the model. For instance, a nonparametric model may contain a latent function which lacks any finite-dimensional representation. This is in contrast with a parametric model where the attention is restricted to a finite-dimensional parameter space. A simple example of a nonparametric model of a PDF is a kernel density estimate. To avoid over-fitting in the nonparametric setting, it is necessary that the model complexity grows in a controlled manner with the amount of data. However, this type of regularization is often intrinsic to the model. Nonparametric models thus avoid the intricate trade-off between model fit and model complexity and at the same time they provide a high degree of flexibility.

In this thesis, we will primarily consider parametric models. Consequently, for clarity of presentation, many of the concepts that we introduce in the sequel are specifically discussed in the context of parametric models. An exception is Paper C, in which a combination of parametric and nonparametric ideas are used to construct a model for a so called Wiener system. The necessary background material on Bayesian nonparametric modeling is given in Section 2.3.

2.2 Maximum likelihood

Consider the parametric model (2.1). Assume that we have collected a batch of data $y_{1:T}$, where T denotes some final time point, i.e. the length of the data record. We refer to the PDF of the measurement sequence $p_\theta(y_{1:T})$ as the *likelihood function*. The likelihood function depends on the model parameter θ. In fact, since the measurement sequence $y_{1:T}$ is assumed to be fixed, it can be viewed as a mapping from the parameter space to the real line,

$$p_\theta(y_{1:T}) : \Theta \to \mathbb{R}. \tag{2.3}$$

A sensible approach to parameter inference is to find a value of θ which maximizes the likelihood function. That is, we seek a parameter value for which the observed data is “as likely as possible”; this idea is known as maximum likelihood (ML). Hence, we define the ML estimator as,

$$\widehat{\theta}_{\text{ML}} = \operatorname*{arg\,max}_{\theta \in \Theta} \log p_\theta(y_{1:T}). \tag{2.4}$$

The logarithm is introduced to simplify and to improve the numerics of the problem. Since the logarithm is strictly increasing, any maximizer of the log-likelihood function is also a maximizer of the likelihood function itself. The ML criterion was proposed, analysed and popularized by Sir Ronald Aylmer Fisher (1890–1962) in the early 20^{th} century (Fisher, 1912, 1921, 1922). However, the idea can be traced back even further to, among others, Gauss, Hagen and Edgeworth (Hald, 1999). Aldrich (1997) provides a historical discussion on Fisher and the making of ML. Due to its appealing theoretical properties, it has a long tradition in many fields of science, including machine learning and system identification.

A challenge in computing the estimator (2.4) for a nonlinear SSM, however, is that the likelihood function in general is not available in closed form. Hence, it is not possible to evaluate the objective function in (2.4), which complicates the optimization problem.

In Section 2.4 we will see how the ML criterion can be related to a state inference problem, which can then be addressed using computational algorithms such as Monte Carlo methods.

2.3 Bayesian learning

An alternative inference strategy bears the name of the British statistician and reverend Thomas Bayes (1702–1761). In Bayesian learning (see e.g. Gelman et al. (2003)), model uncertainties are represented using stochasticity. A probabilistic hypothesis about the model is maintained. When observations regarding the validity of the hypothesis are obtained, the belief in the hypothesis is updated using Bayes' rule. Bayes (1764) treated this problem, but only considered uniform priors. The ideas that we today refer to as Bayesian, were to a large extent pioneered and popularized by the French mathematician Pierre-Simon Laplace (1749–1827). In a memoir, produced at the age of 25 and supposedly unaware of Bayes' work, Laplace (1774) discovered the more general form of Bayes' rule that we use today.

In the parametric setting, the aforementioned hypothesis concerns the model parameters. Consequently, a Bayesian parametric model is characterized by the presence of a prior PDF $\pi(\theta)$ for the model parameter θ, which is thus viewed as a random variable. The prior distribution summarizes our *a priori* knowledge about the parameter, i.e. what we know before we make any observations from the system. Such prior information is sometimes naturally available, e.g. due to physical constraints or insight into the system dynamics based on experience. In other cases, the prior is introduced simply to enable the application of Bayesian methods. In such cases, a pragmatic, but useful, strategy is to choose a prior which results in simple computations. This is achieved by using so called *conjugate priors* (see Section 2.4). It is also common to choose the prior distribution to be uninformative, meaning that it will affect the posterior degree of belief to a small extent.

Given a batch of observations $y_{1:T}$, the Bayesian learning problem amounts to computing the posterior PDF $p(\theta \mid y_{1:T})$. From Bayes' rule, this can be expressed as

$$p(\theta \mid y_{1:T}) = \frac{p_\theta(y_{1:T})\pi(\theta)}{p(y_{1:T})}, \tag{2.5}$$

The above expression relates the posterior PDF to the prior PDF and to the likelihood function. Note that in Bayesian probability θ is viewed as a random variable. Hence, the likelihood function should be thought of as the conditional PDF of the observations given θ, i.e. $p_\theta(y_{1:T}) = p(y_{1:T} \mid \theta)$. However, to be able to discuss the different learning criteria in a common setting, we keep the notation $p_\theta(y_{1:T})$.

If we accept the Bayesian model, the posterior distribution provides a rich source of information about the parameter. It is a complete summary of the *a priori* knowledge and all the information which is available in the observed data. It can for instance be used to compute minimum mean-squared-error estimates of the parameters, but also to systematically reason about the uncertainties in these estimates. Since the posterior PDF depends on the likelihood function, we face similar challenges in computing (2.5) as in solving the ML problem (2.4). We discuss how to make use of computational methods to address this issue in Section 2.4.

In the nonparametric setting, the number of "parameters" is allowed to vary and to grow with the amount of data. Analogously to modeling unknown parameters as latent random variables, Bayesian nonparametric models accomplish this by using latent stochastic processes to represent the unknowns of the model; see e.g. Gershman and Blei (2012); Hjort et al. (2010); Jordan (2010) for an introduction to these models. Principally, learning of Bayesian nonparametric models is similar to learning of parametric models. Using Bayes' rule, the likelihood of the data is combined with the prior to obtain the posterior distribution. However, for a nonparametric model, the target distribution is the posterior law of the latent stochastic process. To illustrate the idea, an example of a Bayesian nonparametric model is given below.

Example 2.1: Gaussian process regression

Regression analysis amounts to learning the relationships within a group of variables. With $\xi \in \mathbb{R}^d$ representing an input variable and $y \in \mathbb{R}$ representing an output variable, we seek a functional relationship such that $y \approx f(\xi)$. The approximate equality reflects the fact that we often observe the function values only up to some uncertainty. Formally, we can write

$$y = f(\xi) + e, \tag{2.6}$$

where e is an error term, here assumed to be zero-mean Gaussian: $e \sim \mathcal{N}(0, \sigma^2)$.

Assume that we observe an input/output data set $\mathcal{D} = \{\xi_i, y_i\}_{i=1}^n$ and wish to estimate the function f. A parametric model can be obtained by fitting, for instance, a polynomial or a trigonometric function to the data. In the nonparametric setting, however, we seek a flexible model where the complexity increases with the number of data points n. One way to accomplish this is to make use of a Gaussian process (GP) regression model (Rasmussen and Williams, 2006).

A GP is a stochastic process, such that any finite collection of sample points have a joint Gaussian distribution. This construction can be used in regression analysis by modeling f (which is indexed by ξ) as a GP with index set $\mathbb{R}^d$. That is, any finite collection of function values, in particular the collection $\{f(\xi_i)\}_{i=1}^n$, have a joint Gaussian distribution. The mean vector and the covariance matrix of this Gaussian distribution follow from the specification of the GP, i.e. from the definition of the prior. Typical choices allow for a high degree of flexibility, but ensure continuity (and sometimes smoothness) of the function f; see Rasmussen and Williams (2006) for a discussion.

Consider now a previously unseen input value $\xi^\star$. From standard manipulations of Gaussian random variables, it follows that the conditional distribution of $f(\xi^\star)$, given $\mathcal{D}$, is also Gaussian with tractable mean and variance. Hence, the posterior GP can be used to predict output values at previously unseen inputs, i.e. it constitutes a model of the function f. The process of GP regression is illustrated in Figure 2.1.

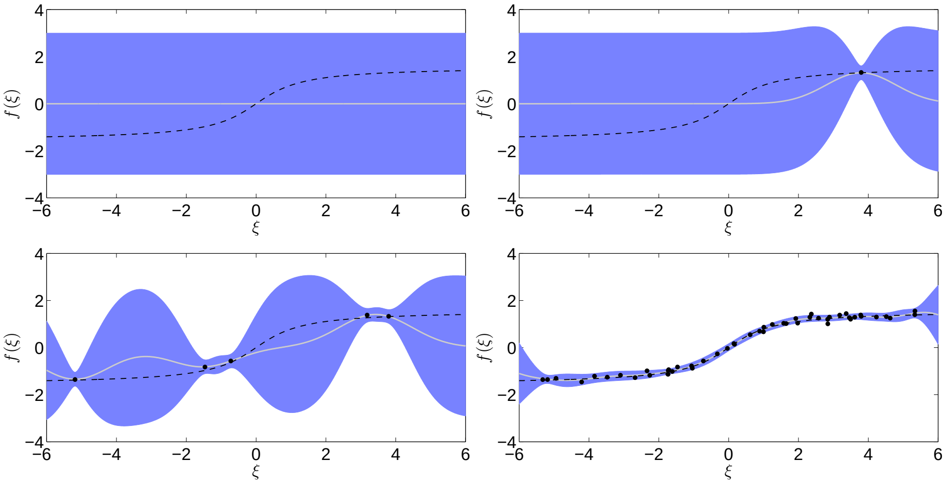

Figure 2.1: *Illustration of GP regression. The mean and the variance ($\pm 3\sigma$) of the GP are show by the solid gray line and by the blue area, respectively. The true, unknown function f is shown by the dashed black line and the data points by black dots. From upper left to lower right; prior GP, i.e. before observing any data, and posterior GP after observing 1, 5 and 50 data points, respectively.*

2.4 Data augmentation

The intractability of the likelihood function appearing in (2.4) and (2.5) is a result of the fact that the state sequence $x_{1:T}$ is latent. Hence, to compute the likelihood of the data, we need to average over all possible state trajectories. More precisely, the likelihood function is given by a marginalization over $x_{1:T}$ according to,

$$p_\theta(y_{1:T}) = \int p_\theta(x_{1:T}, y_{1:T})\, dx_{1:T}. \tag{2.7}$$

Using the conditional independence properties of an SSM, the integrand can be written as,

$$p_\theta(x_{1:T}, y_{1:T}) = \mu_\theta(x_1) \prod_{t=1}^{T} g_\theta(y_t \mid x_t) \prod_{t=1}^{T-1} f_\theta(x_{t+1} \mid x_t). \tag{2.8}$$

The high-dimensional integration in (2.7) will in general lack a closed form solution. This difficulty is central when addressing the learning problem for SSMs. Indeed, the need for using computational methods, such as Monte Carlo, is tightly coupled to the intractability of the above integral. Many of the challenges discussed throughout this thesis is a manifestation of this problem, in one form or another.

The presence of a latent state suggests a technique known as *data augmentation* (Dempster et al., 1977; Tanner and Wong, 1987). While this technique goes beyond learning of SSMs, we discuss how it can be used in our setting below. Data augmentation is based on the idea that if the latent states $x_{1:T}$ would be known, inference about θ would be relatively simple.

This suggests an iterative approach, alternating between updating the belief about $x_{1:T}$ and updating the belief about θ. The former step of the iteration corresponds to solving an intermediate state inference problem. In data augmentation schemes, the states are viewed as *missing data*, as opposed to the *observed data* $y_{1:T}$. That is, the intermediate state inference step amounts to augmenting the observed data, to recover the *complete data* set $\{x_{1:T}, y_{1:T}\}$. The complete data and the observed data likelihoods are related according to (2.7), suggesting that $p_\theta(x_{1:T}, y_{1:T})$ indeed can be useful for drawing inference about θ.

Let us start by considering the Bayesian learning criterion. Assume for the time being that the complete data $\{x_{1:T}, y_{1:T}\}$ is available. From Bayes' rule (cf. (2.5)) we then have,

$$p(\theta \mid x_{1:T}, y_{1:T}) = \frac{p_\theta(x_{1:T}, y_{1:T})\pi(\theta)}{p(x_{1:T}, y_{1:T})}, \tag{2.9}$$

where the complete data likelihood is given by (2.8). While computing the normalization constant in (2.9) can be problematic, it is indeed possible for many models of interest. In particular, for many complete data likelihoods, it is possible to identify a prior PDF $\pi(\theta)$ which is such that the posterior PDF $p(\theta \mid x_{1:T}, y_{1:T})$ belongs to the same family of distributions as the prior. The prior is then said to be *conjugate* to the complete data likelihood (Gelman et al., 2003). For conjugate models, the posterior PDF in (2.9) can be computed in closed form (still, assuming that $x_{1:T}$ is known). All members of the extensive exponential family of distributions have conjugate priors. If the normalization constant cannot be computed in closed form, it is possible to make use of Monte Carlo integration to compute (2.9). We discuss this in more detail in Paper A. See also Paper C, where this technique is used for Wiener system identification.

The problem in using (2.9), however, is that the states $x_{1:T}$ are not known. To address this issue, we will make use of Monte Carlo methods. In particular, one of the main methods that we will consider makes use of the observed data $y_{1:T}$ to impute values for the latent variables $x_{1:T}$ by simulation. Once we have generated a (representative) sample from $x_{1:T}$, this can be used to compute θ according to (2.9). More precisely, we can draw a sample of θ from the posterior distribution (2.9). These two steps are then iterated, i.e. the method alternates between:

(i) Sample $x_{1:T}$ given θ and $y_{1:T}$.

(ii) Sample θ given $x_{1:T}$ and $y_{1:T}$.

This is a so called Gibbs sampler, originating from the method proposed by Geman and Geman (1984). Under appropriate conditions, the distribution of the θ-samples will converge to the target distribution (2.5). Hence, these samples provide an empirical representation of the posterior distribution which is the object of interest in Bayesian learning. The precise way in which the states $x_{1:T}$ are sampled in Step *(i)* will be discussed in detail in Paper A. For now, we note that the Gibbs sampler requires us to generate samples from a, typically, complicated and high-dimensional distribution in order to impute the latent state variables.

Data augmentation is useful also when addressing the ML problem (2.4). Indeed, the technique was popularized in the statistics community by the introduction of the expectation maximization (EM) algorithm by Dempster et al. (1977). EM is a data augmentation algorithm which leverages the idea of missing data to construct a surrogate cost function for the ML problem. Using the relationship

$$p_\theta(x_{1:T} \mid y_{1:T}) = \frac{p_\theta(x_{1:T}, y_{1:T})}{p_\theta(y_{1:T})}, \tag{2.10}$$

the observed data log-likelihood function can be written as

$$\log p_\theta(y_{1:T}) = \log p_\theta(x_{1:T}, y_{1:T}) - \log p_\theta(x_{1:T} \mid y_{1:T}). \tag{2.11}$$

For any $\theta \in \Theta$, $p_\theta(x_{1:T} \mid y_{1:T})$ is a PDF and it thus integrates to one. Hence, by taking an arbitrary $\theta' \in \Theta$, multiplying (2.11) with $p_{\theta'}(x_{1:T} \mid y_{1:T})$ and integrating w.r.t. $x_{1:T}$ we get,

$$\log p_\theta(y_{1:T}) = Q(\theta, \theta') - V(\theta, \theta'), \tag{2.12}$$

where we have defined the auxiliary quantities,

$$\begin{aligned} Q(\theta, \theta') &\triangleq \int \log p_\theta(x_{1:T}, y_{1:T}) p_{\theta'}(x_{1:T} \mid y_{1:T})\, dx_{1:T} \\ &= \mathbb{E}_{\theta'}[\log p_\theta(x_{1:T}, y_{1:T}) \mid y_{1:T}] \end{aligned} \tag{2.13}$$

and $V(\theta, \theta') \triangleq \mathbb{E}_{\theta'}[\log p_\theta(x_{1:T} \mid y_{1:T}) \mid y_{1:T}]$. From (2.12) it follows that, for any $(\theta, \theta') \in \Theta^2$,

$$\log p_\theta(y_{1:T}) - \log p_{\theta'}(y_{1:T}) = (Q(\theta, \theta') - Q(\theta', \theta')) + (V(\theta', \theta') - V(\theta, \theta'))\,. \tag{2.14}$$

The difference $V(\theta', \theta') - V(\theta, \theta')$ can be recognized as the Kullback-Leibler divergence between $p_{\theta'}(x_{1:T} \mid y_{1:T})$ and $p_\theta(x_{1:T} \mid y_{1:T})$, which is known to be nonnegative (Kullback and Leibler, 1951). Hence, as an implication of (2.14) we get,

$$Q(\theta, \theta') \geq Q(\theta', \theta') \Rightarrow \log p_\theta(y_{1:T}) \geq \log p_{\theta'}(y_{1:T}). \tag{2.15}$$

This result implies that the auxiliary quantity (2.13) can be used as a substitute for the log-likelihood function when solving the ML problem (2.4). More precisely, any sequence of iterates which increase the value of the Q-function, will also increase the value of the log-likelihood. This is exploited in the EM algorithm, which iterates between computing the expectation in (2.13) (the E-step) and maximizing the auxiliary quantity $Q(\theta, \theta')$ (the M-step).

The auxiliary quantity of the EM algorithm is defined as the expectation of the complete data log-likelihood according to (2.13). The main challenge in using the EM algorithm for learning of general SSMs lies in the computation of this expectation. However, one possibility is to make use of Monte Carlo methods. That is, we generate samples from the latent states $x_{1:T}$ and approximate the expectation in (2.13) by the sample average. Again, the details of how this simulation can be carried out will be discussed in Paper A.

2.5 Online learning

In the previous sections we have considered *batch-wise* learning. That is, we have assumed that a complete data set $y_{1:T}$, for some final time point T, is available throughout the learning process. In some applications, it is more natural to do the learning *online*, by continuously updating the system model as new observations are obtained (Ljung and Söderström, 1983).

For instance, in the Bayesian, parametric setting, online learning amounts to sequentially computing the posterior PDFs, $p(\theta \mid y_{1:t})$ for $t = 1, 2, \dots$. Similarly, we can construct a sequence of optimization problems as in (2.4) for online ML learning. Online learning is useful in situations where the properties of the system are changing over time. Since the online learning algorithm continuously receive feedback from the system, it can adapt to situations which are previously unseen. Online learning can also be useful in *big data* applications. If the data set is very large, it may be more efficient to process it in an online fashion, i.e. one data item at a time.

We will primarily be concerned with batch-wise learning in this thesis. However, in Paper F, an algorithm for online Bayesian learning of a specific class of SSMs is presented.

3

Monte Carlo methods

This chapter provides an introduction to basic Monte Carlo methods, such as rejection sampling and importance sampling. These are the building blocks for the more advanced methods which are studied throughout the thesis. For a thorough elementary treatment of the Monte Carlo idea, see the books by Robert and Casella (2004) and Liu (2001).

3.1 The Monte Carlo idea

The idea of Monte Carlo methods (see Metropolis and Ulam (1949) for an early discussion) is to make use of random simulation to carry out a computation which is otherwise tedious, or intractable, to perform. For instance, consider the problem of evaluating the intermediate quantity of the EM algorithm in (2.13). This amounts to computing an expected value w.r.t. $p_\theta(x_{1:T} \mid y_{1:T})$, i.e. to solve a generally high-dimensional integration problem. In many cases, and in particular for learning of nonlinear SSMs, this integration lacks a closed form solution. In such situations, Monte Carlo methods can be used to approximate the expected value with a sample average over samples generated from the underlying random variable.

More generally, let $\gamma(x)$ be a PDF, referred to as the target density, which is defined on some space X. Let x be a random variable distributed according to $\gamma(x)$ and assume that we seek the expected value,

$$\mathbb{E}[\varphi(x)] = \int \varphi(x)\gamma(x)\,dx, \tag{3.1}$$

for some test function φ (cf. (2.13)). Let us start by making the assumption that we can generate independent samples $\{x^i\}_{i=1}^N$, distributed according to $\gamma(x)$. This is in fact a very restrictive assumption and a large part of this thesis is concerned with strategies for generating realizations from random variables with complicated distributions. Nevertheless it is instructive to make this assumption in order to be able to focus on the *key idea* underlying all Monte Carlo methods. Based on these samples, we can approximate (3.1) by the sample average,

$$\widehat{\varphi}_{\text{MC}}^{N} \triangleq \frac{1}{N} \sum_{i=1}^{N} \varphi(x^{i}). \tag{3.2}$$

An equivalent interpretation of this Monte Carlo estimator is to let the samples $\{x^i\}_{i=1}^N$ define an empirical approximation of the target distribution,

$$\widehat{\gamma}_{\text{MC}}^{N}(dx) = \frac{1}{N} \sum_{i=1}^{N} \delta_{x^{i}}(dx), \tag{3.3}$$

where $\delta_{x'}(dx)$ denotes a Dirac point-mass located at x'. Hence, we approximate the target distribution (which may be continuous) with a discrete probability distribution, by placing a point-mass probability of $1/N$ at each of the generated samples. Inserting the approximation (3.3) into (3.1) results in

$$\int \varphi(x)\gamma(x)\,dx \approx \int \varphi(x) \frac{1}{N} \sum_{i=1}^{N} \delta_{x^{i}}(dx) = \frac{1}{N} \sum_{i=1}^{N} \varphi(x^{i}), \tag{3.4}$$

i.e. we indeed obtain the Monte Carlo estimator (3.2). The idea of letting a collection of samples define an empirical point-mass distribution as in (3.3) is very convenient and it will be frequently used in the sequel.

The Monte Carlo estimator (3.2) comes with many desirable properties, which to a large extent explains the popularity of the Monte Carlo method. First, it is unbiased, i.e. $\mathbb{E}[\widehat{\varphi}_{\text{MC}}^{N}] = \mathbb{E}[\varphi(x)]$ where the former expectation is w.r.t. the random realizations $\{x^i\}_{i=1}^N$. Second, the strong law of large numbers implies almost sure convergence, $\widehat{\varphi}_{\text{MC}}^{N} \xrightarrow{\text{a.s.}} \mathbb{E}[\varphi(x)]$ as $N \to \infty$. Additionally, if the variance of $\varphi(x)$ is finite, i.e. $\sigma_\varphi^2 = \text{Var}[\varphi(x)] < \infty$, then a central limit theorem (CLT) holds,

$$\frac{\sqrt{N}\left(\widehat{\varphi}_{\text{MC}}^{N} - \mathbb{E}[\varphi(x)]\right)}{\sigma_\varphi} \xrightarrow{\text{D}} \mathcal{N}(0,1), \qquad N \to \infty, \tag{3.5}$$

where $\xrightarrow{\text{D}}$ denotes convergence in distribution. In fact, the variance of the estimator (3.2) is explicitly given by $\text{Var}[\widehat{\varphi}_{\text{MC}}^{N}] = \sigma_\varphi^2/N$. From (3.5) if follows that the Monte Carlo error decreases as $O(N^{-\frac{1}{2}})$. Interestingly, the convergence rate is independent of the dimension of X. This clearly distinguishes Monte Carlo methods from deterministic integration methods, where the latter have an approximation error that grows with the dimension.

As pointed out above, the vanilla Monte Carlo method described above relies on the, often unrealistic, assumption that it is possible to generate independent and identically distributed (i.i.d.) samples from the target distribution. The rest of this chapter, and to a large extent Part II of this thesis, is devoted to strategies for generating samples from complicated target distributions, effectively rendering the use of Monte Carlo methods possible for challenging inference and learning problems.

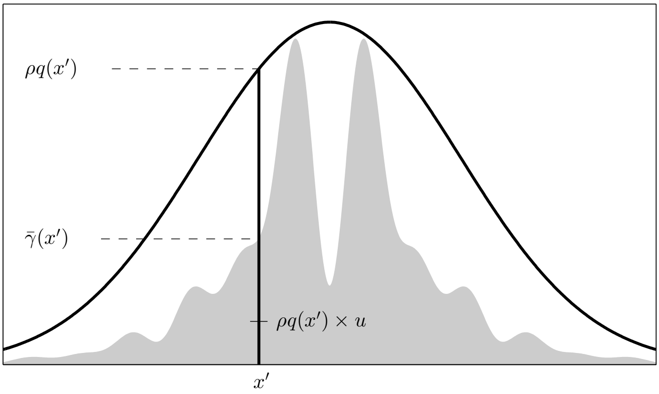

Figure 3.1: *Illustration of rejection sampling. The graph of $\bar{\gamma}(x)$ (gray area) is bounded by the graph of $\rho q(x)$ (black curve). A sample is generated uniformly over the area under the black curve. If the sample falls is the gray area, it is accepted as a draw from $\gamma(x)$, otherwise it is rejected.*

3.2 Rejection Sampling

An often encountered difficulty is that the target density $\gamma(x)$ can be evaluated only up to proportionality. That is, we can write $\gamma(x) = \bar{\gamma}(x)/Z$, where $\bar{\gamma}(x)$ can be evaluated point-wise, but where the normalization constant Z is unknown. The typical setting is when Bayes' rule is used to express a posterior PDF in terms of the prior, the likelihood and the (unknown) normalizing constant. For instance, consider the Bayesian learning criteria (2.5). Even in situations when the likelihood function is available, if the model is non-conjugate, then the normalization constant $p(y_{1:T})$ is typically unknown.

Rejection sampling (von Neumann, 1951) is a Monte Carlo method which, under these conditions, can be used to generate samples *exactly* distributed according to the target density $\gamma(x)$. To introduce the idea, let $\bar{\gamma}(z)$ be given by the function shown by the gray area in Figure 3.1. Let the two-dimensional random vector (x, y) be distributed uniformly over the gray area. The area under the graph of $\bar{\gamma}(x)$ is $\int \bar{\gamma}(x)\, dx = Z$ which implies that the PDF of (x, y) is,

$$p(x, y) = \begin{cases} 1/Z & \text{if } 0 \leq y \leq \bar{\gamma}(x), \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

Hence, the marginal PDF of x is,

$$p(x) = \int p(x, y)\, dy = \int_0^{\bar{\gamma}(x)} \frac{1}{Z}\, dy = \gamma(x), \tag{3.7}$$

i.e. it holds that x is marginally distributed according to the target distribution.

The problem is that sampling uniformly over the gray area is just as hard as the original problem, i.e. sampling from $\gamma(z)$. However, it leads us to the following idea. Let $q(x)$

Algorithm 1 Rejection sampling

```
$L \leftarrow \{1, \ldots, N\}$.
while $L$ is not empty do
    $n \leftarrow \mathrm{card}(L)$.
    $\delta \leftarrow \emptyset$.
    Sample independently $\{x'(k)\}_{k=1}^{n} \sim q(x)$.
    Sample independently $\{u(k)\}_{k=1}^{n} \sim \mathcal{U}([0,1])$.
    for $k = 1$ to $n$ do
        if $u(k) \leq \frac{\bar{\gamma}(x'(k))}{\rho q(x'(k))}$ then
            $x^{L(k)} \leftarrow x'(k)$.
            $\delta \leftarrow \delta \cup \{L(k)\}$.
        end if
    end for
    $L \leftarrow L \setminus \delta$.
end while
```

be a user-chosen PDF which is easy to sample from. Such a distribution is referred to as a *proposal* distribution. Furthermore, assume that there exists a constant ρ such that $\bar{\gamma}(x) \leq \rho q(x)$ for all $x \in \mathsf{X}$. Now, if we sample independently and uniformly under the graph of $\rho q(x)$, but only keep the samples that fall under the graph of $\bar{\gamma}(x)$, then the surviving samples are i.i.d. draws from the target distribution.

More generally, let x' be sampled from the proposal and let u be drawn uniformly over the unit interval, i.e.

$$x' \sim q(x), \tag{3.8a}$$

$$u \sim \mathcal{U}([0,1]). \tag{3.8b}$$

The variable u serves as an indicator on whether we should accept x' as a valid sample from the target distribution or not. More precisely, if $\rho q(x')u \leq \bar{\gamma}(x')$ we set $x = x'$, otherwise we reject x' and repeat the procedure (3.8) until a sample is accepted. The method is summarized in Algorithm 1, in which N i.i.d. samples $\{x^i\}_{i=1}^{N}$ are generated in parallel.

To see the validity of the algorithm, consider the probability that x falls in some subset $A \subset \mathsf{X}$,

$$\mathbb{P}(x \in A) = \mathbb{P}(x' \in A \mid x' \text{ is accepted}) = \frac{\mathbb{P}(x' \in A \cap x' \text{ is accepted})}{\mathbb{P}(x' \text{ is accepted})}. \tag{3.9}$$

Since x' is distributed according to $q(x)$ and u is uniform on $[0,1]$, the numerator can be expressed as

$$\mathbb{P}(x' \in A \cap u \leq \bar{\gamma}(x')/(\rho q(x'))) = \int_A \frac{\bar{\gamma}(x')}{\rho q(x')} q(x')\, dx' = \frac{Z}{\rho} \int_A \gamma(x')\, dx'. \tag{3.10}$$

Analogously, the denominator in (3.9) is given by $\mathbb{P}(x' \text{ is accepted}) = Z/\rho$. Inserting this into (3.9) we get,

$$\mathbb{P}(x \in A) = \int_A \gamma(x)\, dx. \tag{3.11}$$

Since the set A is arbitrary, we conclude that x is indeed distributed according to the target density $\gamma(x)$.

The choice of proposal density $q(x)$ and the constant ρ are very important from a practical point of view. As noted above, the *acceptance probability* is given by $\mathbb{P}(x' \text{ is accepted}) = Z/\rho$. Consequently, the average number of candidate samples that need to be drawn to generate one sample from the target distribution is ρ/Z. It is therefore imperative that this ratio is not too large. However, for the algorithm to work, we also require that the graph of $\bar{\gamma}(x)$ is completely below the graph of $\rho q(x)$, i.e. ρ is at least as large as the largest discrepancy between the proposal $q(x)$ and the (unnormalized) target $\bar{\gamma}(x)$. Finding a proposal density in close agreement with the target density is easy for the toy problem considered above. However, as the target density becomes more complicated and, in particular, as the dimension of X increases, this becomes harder.

For the sake of illustration, assume that we wish sample from the d-dimensional, standard normal distribution using rejection sampling. As proposal, we use a d-dimensional, zero-mean normal distribution with covariance matrix $\sigma_q^2 I_d$. For the ratio between the target and the proposal densities to be bounded, we require that $\sigma_q \geq 1$. The smallest bound is then given by $\rho = \sigma_q^d$. Hence, the acceptance probability decays exponentially as we increase the dimension of the problem. This is referred to as the *curse of dimensionality*. In high dimensions, what appears to be a small discrepancy between the proposal and the target densities, can in fact have a huge impact, rendering the method impractical.

3.3 Importance sampling

Importance sampling (Kahn and Harris, 1951; Marshall, 1956) offers a solution to the problem of evaluating integrals of the form (3.1), but it does not generate exact draws from the target distribution. In the rejection sampler introduced above, we first generate candidate samples from some proposal density $q(x)$. These samples are then either accepted or rejected with certain probabilities, depending on how well they fit the target distribution. Importance sampling proceeds similarly, by generating draws from a proposal distribution. However, rather than discarding some of the simulated values all samples are kept, but they are assigned individual weights depending on how well they fit the target.

Let $x' \sim q(x)$ be an instrumental random variable, distributed according to the proposal. We can then express (3.1) as,

$$\begin{aligned}\mathbb{E}[\varphi(x)] &= \int \varphi(x)\gamma(x)\, dx = \int \varphi(x)\frac{\gamma(x)}{q(x)}q(x)\, dx \\ &= \int \varphi(x)W(x)q(x)\, dx = \mathbb{E}[\varphi(x')W(x')],\end{aligned} \tag{3.12}$$

where we have introduced the *weight function* $W(x) \triangleq \gamma(x)/q(x)$ and where we have assumed that $q(x) > 0$ for all x where $\varphi(x)\gamma(x) > 0$ (i.e. $\operatorname{supp} \varphi\gamma \subset \operatorname{supp} q$). By construction, it is easy to generate samples from $q(x)$. We can thus construct a Monte Carlo estimator for (3.12) by sampling independently $x^i \sim q(x)$ for $i = 1, \ldots, N$ and computing,

$$\bar{\varphi}_{\mathrm{IS}}^N \triangleq \frac{1}{N} \sum_{i=1}^{N} W(x^i)\varphi(x^i). \tag{3.13}$$

This estimator is similar to (3.2), but we see that the samples are weighted by so called *importance weights*, accounting for the discrepancy between the proposal and the target densities. Intuitively speaking, the importance weights contain information about how useful each proposed value x^i is for computing integrals on the form (3.1).

As mentioned in the previous section, it is common that it is only possible to evaluate the target density up to an unknown normalization constant. That is, we can write $\gamma(x) = \bar{\gamma}(x)/Z$, where $\bar{\gamma}(x)$ can be evaluated but the constant Z is unknown[1]. We then have,

$$\mathbb{E}[\varphi(x)] = \int \varphi(x) \frac{\bar{\gamma}(x)}{Zq(x)} q(x)\, dx = \frac{1}{Z} \int \varphi(x) W(x) q(x)\, dx, \tag{3.14}$$

where (with abuse of notation) we have redefined the weight function as

$$W(x) \triangleq \frac{\bar{\gamma}(x)}{q(x)}. \tag{3.15}$$

Hence, the importance sampling estimator (3.13) is given by,

$$\bar{\varphi}_{\mathrm{IS}}^N = \frac{1}{NZ} \sum_{i=1}^{N} \bar{w}^i \varphi(x^i), \tag{3.16}$$

where we have explicitly introduced the weights $\bar{w}^i = W(x^i)$ for $i = 1, \ldots, N$. Note that, since $\bar{w}^i$ is given by a transformation of a random variable, it is itself a random variable. From the above expression it appears as if we have just moved the problem with the unknown normalization constants from one place to another. However, we can make use of the samples $\{x^i\}_{i=1}^N$ to compute an approximation of the unknown constant. Indeed, the normalization constant is given by,

$$Z = \int \bar{\gamma}(x)\, dx = \int \frac{\bar{\gamma}(x)}{q(x)} q(x)\, dx \approx \frac{1}{N} \sum_{i=1}^{N} \bar{w}^i. \tag{3.17}$$

By inserting this approximation into (3.16), we obtain the *normalized* importance sampling estimator,

$$\widehat{\varphi}_{\mathrm{IS}}^N = \sum_{i=1}^{N} w^i \varphi(x^i), \tag{3.18}$$

where $\{w^i\}_{i=1}^N$ denote the normalized importance weights: $w^i \triangleq \bar{w}^i / \sum_l \bar{w}^l$. Analogously to (3.3), an alternative interpretation of the above is that the importance sampler provides an empirical point-mass approximation of the target distribution, according to,

---

[1] Similarly, we may assume that the proposal density can only be evaluated up to proportionality, but that is less common in practice.

$$\widehat{\gamma}_{\text{IS}}^{N}(dx) = \sum_{i=1}^{N} w^i \delta_{x^i}(dx). \tag{3.19}$$

Hence, even though the importance sampler does *not* provide samples from the target distribution, the weighted samples $\{x^i, w^i\}_{i=1}^N$ define an empirical distribution approximating the target. Inserting this empirical distribution into (3.1) straightforwardly results in the estimator (3.18). Note that, even if the constant Z is known, the importance weights must be normalized for the point-mass approximation (3.19) to be a probability distribution. The above development is summarized in Algorithm 2.

Algorithm 2 Importance sampling (all operations are for $i = 1, \dots, N$)

```
Draw $x^i \sim q(x)$.
Compute $\bar{w}^i = W(x^i)$.
Normalize: set $w_1^i = \bar{w}_1^i / \sum_l \bar{w}^l$.
```

From the discussion on the rejection sampler in Section 3.2, we recall that the choice of proposal distribution is important in order to obtain a practical algorithm. This holds true also for the importance sampler. A large discrepancy between the target and the proposal densities will lead to high variance in the importance weights, which carries over to the estimator (3.18).

3.4 Particle filters and Markov chains

Rejection sampling and importance sampling are important tools in the construction of Monte Carlo inferential methods. However, alone they do not provide satisfactory solutions to the challenging inference problems associated with learning of dynamical systems. When dealing with SSMs, data augmentation schemes typically introduce the state sequence $x_{1:T}$ as auxiliary variables (see Section 2.4). Hence, the dimensionality of the problem increases with the length of the data record T, which is typically large. For these learning problems, rejection sampling and importance sampling algorithms are often impractical, due to the difficulty of designing efficient proposal distributions in high dimensions. Consequently, there is a need for Monte Carlo methods which are more apt at addressing high-dimensional integration problems.

Two classes of algorithms play a central role in this thesis and, indeed, in the study of Monte Carlo methods as a whole. These are, respectively, methods based on interacting particle systems and on Markov chains. However, despite the relevance of these algorithms for this thesis, this background section only provides a very brief introduction to the concepts. The reason is that both classes of algorithms are introduced and more thoroughly described in Section 1 of Paper A.

The former class of methods are referred to as particle filters or sequential Monte Carlo (SMC) methods (Stewart and McCarty, 1992; Gordon et al., 1993); see also Doucet and

Johansen (2011); Gustafsson (2010); Doucet et al. (2001). Particle filters are useful for approximating a sequence of target distributions. For instance, in the context of SSMs, it is common to target the sequence of *joint smoothing densities* $p_\theta(x_{1:t} \mid y_{1:t})$ for $t = 1, 2, \ldots$. Initially, for $t = 1$, the density $p_\theta(x_1 \mid y_1)$ is approximated using importance sampling. This is much simpler than targeting, say, $p_\theta(x_{1:T} \mid y_{1:T})$ directly, due to the high dimensionality of the latter density for large T. Hence, we obtain an empirical point-mass approximation as in (3.19);

$$\widehat{p}_\theta^N(dx_1 \mid y_1) = \sum_{i=1}^{N} w_1^i \delta_{x_1^i}(dx_1). \tag{3.20}$$

In the SMC literature, the samples $\{x_1^i\}_{i=1}^N$ are called *particles* and $\{x_1^i, w_1^i\}_{i=1}^N$ is referred to as a *weighted particle system*. The particles can be thought of as (random) hypotheses about the state of the system at time $t = 1$. The belief in each of the hypotheses is represented by the corresponding importance weight.

The essence of the particle filter is a systematic procedure for updating these hypotheses to obtain an approximation of the next target density in the sequence. That is, given (3.20) we seek a point-mass approximation of $p_\theta(x_{1:2} \mid y_{1:2})$, then of $p_\theta(x_{1:3} \mid y_{1:3})$, and so on. Basically, this is accomplished by propagating and reevaluating the belief in the particles according to the model (2.1). By discarding or duplicating particles according to their importance weights, the particle filter is able to put emphasis on high-probability hypotheses, which are more likely to be useful for approximating the next target distribution in the sequence.

The second class of methods are so called Markov chain Monte Carlo (MCMC) samplers. A Markov chain is a memoryless stochastic process. That is, the next state of the chain depends only on the current state and not on the past history of the process. In MCMC, Markov chains are used to represent a sequence of hypotheses about some variable of interest, e.g. the state of a dynamical system at some specific time point or some unknown model parameter. MCMC samplers are thus iterative Monte Carlo methods where each sample (i.e. each hypothesis) is statistically dependent on the previous one. For this approach to be useful for inference, the Markov chain has to be constructed in such a way that, in the long run, the samples are representative for the target distribution. That is, the *limiting distribution* of the chain should coincide with the target distribution. In MCMC theory, there are systematic ways of constructing Markov chains with this specific property. The Metropolis-Hastings sampler (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984) are the most well-known techniques. The latter method was briefly mentioned in the context of data augmentation in Section 2.4,

It is also possible to combine SMC and MCMC to construct composite algorithms, drawing on the strengths of both classes of methods. This thesis puts particular emphasis on a class of methods referred to a particle MCMC (PMCMC) algorithms (Andrieu et al., 2010). PMCMC relies on SMC for generating samples of the often highly auto-correlated state trajectory. Combined with MCMC, this results in powerful Monte Carlo methods which are capable of addressing both inference and learning problems in complex dynamical systems.

3.5 Rao-Blackwellization

In the mid 40's, Rao (1945) and Blackwell (1947) established a fundamental result in estimation theory, which has later become known as the Rao-Blackwell theorem (see also (Lehmann, 1983, page 50)). Let θ be an unknown parameter and let Y be some data drawn from a distribution parameterized by θ. Given the data Y, we compute an estimator of θ denoted as $\widehat{\theta}(Y)$. Furthermore, let S be a sufficient statistic for Y, i.e. informally S contains the same amount of information about θ as Y does. Then, basically, the Rao-Blackwell theorem states that

$$\widehat{\theta}_{\text{RB}}(S) = \mathbb{E}[\widehat{\theta}(Y) \mid S] \tag{3.21}$$

is typically a better estimator than $\widehat{\theta}(Y)$, and it is never worse. Hence, from a crude estimator $\widehat{\theta}(Y)$ we can construct a better estimator $\widehat{\theta}_{\text{RB}}(S)$, depending only on the sufficient statistic S, by computing a conditional expectation. This transformation is known as *Rao-Blackwellization.*

In this thesis, we are concerned with the implication of the Rao-Blackwell theorem for estimators constructed using Monte Carlo methods. Any Monte Carlo estimator is affected by variance due to the random simulation used in its construction. For instance, consider the vanilla Monte Carlo estimator in (3.2),

$$\widehat{\varphi}^N_{\text{MC}}(X) = \frac{1}{N}\sum_{i=1}^{N} \varphi(x^i). \tag{3.22}$$

Here, we have explicitly introduced the dependence on the random samples $X \triangleq \{x^i\}_{i=1}^N$ in the notation. As previously mentioned, if the variance of the test function $\varphi(x)$ is finite, we have

$$\text{Var}(\widehat{\varphi}^N_{\text{MC}}(X)) = \frac{\text{Var}(\varphi(x))}{N}. \tag{3.23}$$

This Monte Carlo variance reveals that there is a random error in the estimator. In fact, since (3.22) is unbiased the mean-squared-error is given by the variance (3.23). Hence, to obtain an accurate estimator it is desirable to keep the variance as small as possible.

By making use of Rao-Blackwellization, it is possible to reduce the Monte Carlo variance and thereby improve upon (3.22). Let the random vector x be split into two components,

$$x = \begin{pmatrix} z \\ \xi \end{pmatrix}. \tag{3.24}$$

The samples $\{x^i\}_{i=1}^N$ are split accordingly and we can thus identify $X = \{Z, \Xi\}$, where $Z = \{z^i\}_{i=1}^N$ and $\Xi = \{\xi^i\}_{i=1}^N$. Furthermore, the test function can be written as $\varphi(x) = \varphi(z, \xi)$ and it follows the estimator (3.22) can be written,

$$\widehat{\varphi}^N_{\text{MC}}(Z, \Xi) = \frac{1}{N}\sum_{i=1}^{N} \varphi(z^i, \xi^i). \tag{3.25}$$

Now, consider the Rao-Blackwellized estimator,

$$\widehat{\varphi}^N_{\text{RB}}(\Xi) = \mathbb{E}[\widehat{\varphi}^N_{\text{MC}}(Z, \Xi) \mid \Xi] = \frac{1}{N}\sum_{i=1}^{N} \varphi^c(\xi^i), \tag{3.26}$$

where we have introduced the function $\varphi^c(\xi) = \mathbb{E}[\varphi(z, \xi) \mid \xi]$. Note that (3.26) depends only on Ξ. Similarly to (3.23), the variance of the estimator (3.26) is given by

$$\text{Var}(\widehat{\varphi}^N_{\text{RB}}(\Xi)) = \frac{\text{Var}(\varphi^c(\xi))}{N}. \tag{3.27}$$

From the law of total variance, it follows that

$$\text{Var}(\varphi(z, \xi)) = \text{Var}(\mathbb{E}[\varphi(z, \xi) \mid \xi]) + \underbrace{\mathbb{E}[\text{Var}(\varphi(z, \xi) \mid \xi)]}_{\geq 0}. \tag{3.28}$$

The term on the left-hand-side corresponds to the variance of the vanilla Monte Carlo estimator (3.22) and the first term on the right-hand-side corresponds to the variance of the Rao-Blackwellized estimator (3.26). Since the second term on the right side is non-negative, it follows that (3.26) dominates (3.22), as claimed.

For the Rao-Blackwellized estimator (3.26) to be practical, it is necessary that the conditional expectation,

$$\varphi^c(\xi) = \mathbb{E}[\varphi(z, \xi) \mid \xi] = \int \varphi(z, \xi)\gamma(z \mid \xi)\, dz, \tag{3.29}$$

can be computed efficiently. Hence, Rao-Blackwellization is useful when, in some sense, part of the integration problem in (3.1) is analytically tractable. In fact, if we set $Z = X$ (and thus $\Xi = \emptyset$), the Rao-Blackwellized estimator (3.26) is given by $\widehat{\varphi}^N_{\text{RB}} = \mathbb{E}[\varphi(x)]$. Clearly, if it is intractable to solve the integration problem in (3.1) in the first place, then computing this "fully Rao-Blackwellized" estimator is also intractable (since they coincide). Hence, there is a trade-off between using Monte Carlo methods to construct randomized estimators, and the application of Rao-Blackwellization to these estimators. The general idea that will be applied in this thesis is to make use of Rao-Blackwellization to an as large degree as possible; see Papers E and F.

In the context of SMC, Rao-Blackwellization is often used to analytically marginalize over part of the state-space. That is, similarly to (3.24), the state is split into two components, $x_t = (z_t, \xi_t)$. The component ξ_t is represented using particles, whereas z_t is marginalized. This results in the so called Rao-Blackwellized particle filter (RBPF) (Chen and Liu, 2000; Doucet et al., 2000; Schön et al., 2005). The most well-known application of the RBPF is for a class of SSMs in which z_t is conditionally linear Gaussian. The aforementioned marginalization then amounts to running a conditional Kalman filter (Kalman, 1960) for each particle to marginalize z_t.

For the vanilla Monte Carlo method, the variance reduction offered by Rao-Blackwellization is straightforwardly quantified by considering a decomposition of variance as in (3.28). This analysis, however, does not apply to more advanced Rao-Blackwellized Monte Carlo methods, such as the RBPF. This issue is addressed in Paper G, where the RBPF is analysed and a variance reduction expression akin to (3.28) is given.

4

Concluding remarks

This chapter concludes the first part of the thesis and points out possible directions for future work. Note, however, that more detailed discussions can be found in the concluding sections of the individual papers in Part II of the thesis.

4.1 Conclusions and future work

The contributions of the thesis can be grouped into two categories – those based on PGAS and those based on Rao-Blackwellization.

The PGAS algorithm has been found to be a useful tool for a range of tasks related to inference and learning of dynamical systems. Various algorithms have been developed around PGAS, addressing both Bayesian and maximum-likelihood-based learning as well as state inference. It has been illustrated how PGAS can be used to solve the classical problem of Wiener system identification in a challenging setting. Basic convergence results have been obtained for PGAS. However, it remains a topic for future work to establish stronger and more explicit ergodicity results, providing informative rates of convergence of the algorithm.

Another direction for future work is to adapt PGAS to certain model classes for which the basic algorithm is not applicable. PGAS relies heavily on so called *ancestor sampling*. To implement this procedure, it is necessary to evaluate point-wise the transition density function of the model under study (see Papers A and B for details). However, for many models encountered in practice this is not possible. Indeed, the transition density function may not even exist! One option is to study these specific models in more detail and thereby try to modify the PGAS algorithm so that it becomes applicable in these scenarios.

Another possibility, however, is to note that many of the models for which ancestor sampling is problematic can be reformulated as non-Markovian latent variable models (see Section 4.6 of Paper A). This is interesting, since we have shown that PGAS can be useful for inference and learning of precisely such non-Markovian models. Hence, this opens up for using PGAS also for the aforementioned model classes, for which ancestor sampling is not directly possible. This is an encouraging result, since there has not been much progress made in solving inference and learning problems for non-Markovian models. However, it remains to evaluate and to better understand the properties of the PGAS method when applied to these models.

There are, of course, other limitations of the PGAS method as well. In particular: *(i)* the method may converge slowly when there are strong dependencies between the states and the parameters of the model, and *(ii)* the method can only be used for batch data (i.e. offline). One interesting direction of future work is to investigate possible ways in which these limitations can be mitigated.

The contributions of the thesis which are based on Rao-Blackwellization include the development of a Rao-Blackwellized particle smoother (RBPS) and a method for online Bayesian parameter estimation. Furthermore, the asymptotic variance of the RBPF has been analysed and compared to that of the standard particle filter. An interesting topic for future work is to establish a similar variance reduction result for the proposed RBPS, i.e. to answer the question of how much we benefit from Rao-Blackwellization when addressing the smoothing problem.

The proposed method for online Bayesian parameter estimation is based on the RBPF. The algorithm can be used for identification of nonlinear SSMs with an affine dependence on the parameters. A Gaussian mixture representation of the posterior parameter distribution $p(\theta \mid y_{1:t})$ is maintained. To mitigate the so called path degeneracy problem, which prevents accurate learning of the model parameters using a standard RBPF, a *mixing step* is incorporated in the algorithm. Unfortunately, this gives rise to a computational complexity which scales quadratically with the number of particles. Future work is needed in order to obtain a computationally more efficient algorithm. Furthermore, it would be interesting to modify the algorithm so that it can be used also for non-affine models. In particular, it should be possible to use the same idea for any model belonging to the exponential family.

Finally, one direction of future work which applies to all the methods discussed throughout the thesis, is to evaluate them in real applications. So far, the methods have primarily been tested in simulation studies. While this is of course very useful as a first level of evaluation, it is not until the methods are used to solve real and relevant problems that their true values can be determined.

4.2 Further reading

The learning problem is discussed in a general setting in the textbooks by Hastie et al. (2009) and Barber (2012). The latter contains one part dedicated to dynamical systems. Ljung (1999); Söderström and Stoica (1989) provide a detailed coverage of the problem from a system identification point of view. For readers who are interested in learning more about Monte Carlo methods, the books by Robert and Casella (2004); Liu (2001) provide a thorough introduction. MCMC and SMC methods are discussed in detail in the collections by Brooks et al. (2011) and Doucet et al. (2001), respectively. See also Part VII of (Crisan and Rozovskii, 2011). Del Moral (2004) provides a extensive collection of convergence results for SMC. Finally, the textbooks by Cappé et al. (2005); Schön and Lindsten (2013) focus on using Monte Carlo methods for inference and learning of dynamical systems.

Bibliography

J. Aldrich. R. A. Fisher and the making of maximum likelihood 1912–1922. *Statistical Science*, 12(3):162–176, 1997.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 72(3):269–342, 2010.

D. Barber. *Bayesian Reasoning and Machine Learning.* Cambridge University Press, 2012.

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1764.

D. Blackwell. Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, 18(1):105–110, 1947.

S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors. *Handbook of Markov Chain Monte Carlo.* Chapman & Hall/CRC, 2011.

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

R. Chen and J. S. Liu. Mixture Kalman filters. *Journal of the Royal Statistical Society: Series B*, 62(3):493–508, 2000.

D. Crisan and B. Rozovskii, editors. *The Oxford Handbook of Nonlinear Filtering.* Oxford University Press, 2011.

J. Dahlin, F. Lindsten, T. B. Schön, and A. Wills. Hierarchical Bayesian ARX models for robust inference. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, July 2012.

J. Dahlin, F. Lindsten, and T. B. Schön. Particle Metropolis Hastings using Langevin dynamics. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

P. Del Moral. *Feynman-Kac Formulae - Genealogical and Interacting Particle Systems with Applications.* Probability and its Applications. Springer, 2004.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

A. Doucet and A. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan and B. Rozovskii, editors, *The Oxford Handbook of Nonlinear Filtering*. Oxford University Press, 2011.

A. Doucet, S. J. 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, editors. *Sequential Monte Carlo Methods in Practice*. Springer Verlag, New York, USA, 2001.

A. Eskandarian, editor. *Handbook of Intelligent Vehicles*. Springer, 2012.

R. A. Fisher. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–160, 1912.

R. A. Fisher. On the“probable error” of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London Series A*, 222:309–368, 1922.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2003.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

S. J. Gershman and D. M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56:1–12, 2012.

J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107 –113, April 1993.

W. H. Greene. *Econometric Analysis*. Prentice Hall, 7th edition, 2008.

F. Gustafsson. Particle filter theory and practice with positioning applications. *IEEE Aerospace and Electronic Systems Magazine*, 25(7):53–82, 2010.

A. Hald. On the history of maximum likelihood in relation to inverse probability and least squares. *Statistical Science*, 14(2):214–222, 1999.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors. *Bayesian Nonparametrics*. Cambridge University Press, 2010.

J. C. Hull. *Options, Futures, and Other Derivatives*. Prentice Hall, 8th edition, 2011.

M. I. Jordan. Bayesian nonparametric learning: Expressive priors for intelligent systems. In R. Dechter, H. Geffner, and J. Halpern, editors, *Heuristics, Probability and Causality: A Tribute to Judea Pearl*. College Publications, 2010.

H. Kahn and T. E. Harris. Estimation of particle transmission by random sampling. *Journal of Research of the National Bureau of Standards, Applied Mathematics Series*, 12: 27–30, 1951.

T. Kailath. *Linear Systems*. Prentice Hall, New Jersey, USA, 1980.

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

M.J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2007.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

P. S. Laplace. Mémoire sur la probabilité des causes par les évènemens. *Mémoires de mathématique et de physique presentés á l'Académie royale des sciences par divers savants & lus dans ses assemblées*, 6:621–656, 1774.

E. L. Lehmann. *Theory of Point Estimation*. Probability and mathematical statistics. John Wiley & Sons, New York, USA, 1983.

F. Lindsten. An efficient stochastic approximation EM algorithm using conditional particle filters. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

F. Lindsten and T. B. Schön. Identification of mixed linear/nonlinear state-space models. In *Proceedings of the 49th IEEE Conference on Decision and Control (CDC)*, Atlanta, USA, December 2010.

F. Lindsten and T. B. Schön. On the use of backward simulation in the particle Gibbs sampler. In *Proceedings of the 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012.

F. Lindsten and T. B. Schön. On the use of backward simulation in particle Markov chain Monte Carlo methods. arXiv.org, arXiv:1110.2873v2, March 2012.

F. Lindsten and T. B. Schön. Backward simulation methods for Monte Carlo statistical inference. *Foundations and Trends in Machine Learning*, 6(1):1–143, 2013.

F. Lindsten, P.-J. Nordlund, and F. Gustafsson. Conflict detection metrics for aircraft sense and avoid systems. In *Proceedings of the 7th IFAC Symposium on Fault Detection,* *Supervision and Safety of Technical Processes (SafeProcess)*, Barcelona, Spain, July 2009.

F. Lindsten, J. Callmer, H. Ohlsson, D. Törnqvist, T. B. Schön, and F. Gustafsson. Georeferencing for UAV navigation using environmental classification. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Anchorage, USA, May 2010.

F. Lindsten, H. Ohlsson, and L. Ljung. Clustering using sum-of-norms regularization; with application to particle filter output computation. In *Proceedings of the IEEE Workshop on Statistical Signal Processing (SSP)*, Nice, France, June 2011a.

F. Lindsten, T. B. Schön, and J. Olsson. An explicit variance reduction expression for the Rao-Blackwellised particle filter. In *Proceedings of the 18th IFAC World Congress*, Milan, Italy, August 2011b.

F. Lindsten, M. I. Jordan, and T. B. Schön. Ancestor sampling for particle Gibbs. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS) 25*, pages 2600–2608. 2012a.

F. Lindsten, T. B. Schön, and M. I. Jordan. A semiparametric Bayesian approach to Wiener system identification. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, July 2012b.

F. Lindsten, T. B. Schön, and L. Svensson. A non-degenerate Rao-Blackwellised particle filter for estimating static parameters in dynamical models. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, July 2012c.

F. Lindsten, P. Bunch, S. J. Godsill, and T. B. Schön. Rao-Blackwellized particle smoothers for mixed linear/nonlinear state-space models. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013a.

F. Lindsten, T. B. Schön, and M. I. Jordan. Bayesian semiparametric Wiener system identification. *Automatica*, 49(7):2053–2063, 2013b.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, 2001.

L. Ljung. *System identification, Theory for the user.* System sciences series. Prentice Hall, Upper Saddle River, NJ, USA, second edition, 1999.

L. Ljung and T. Söderström. *Theory and Practice of Recursive Identification.* MIT Press, 1983.

A. W. Marshall. The use of multistage sampling schemes in Monte Carlo computations. In H. A. Meyer, editor, *Symposium on Monte Carlo Methods.* John Wiley & Sons, New York, USA, 1956.

N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6): 1087–1092, 1953.

A. V. Oppenheim, A. S. Willsky, and S. H. Nawab. *Signals and Systems*. Prentice Hall, 2nd edition, 1996.

C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

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

D. A. Rasmussen, O. Ratmann, and K. Koelle. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biology*, 7(8), 2011.

K. J. Åström and R. M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2008.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

T. Schön, F. Gustafsson, and P.-J. Nordlund. Marginalized particle filters for mixed linear/nonlinear state-space models. *IEEE Transactions on Signal Processing*, 53(7):2279–2289, July 2005.

T. B. Schön and F. Lindsten. *Computational Learning in Dynamical Systems*. 2013. (forthcoming, draft manuscript is available from the authors).

T. Söderström and P. Stoica. *System Identification*. Prentice Hall, 1989.

L. Stewart and P. McCarty. The use of Bayesian belief networks to fuse continuous and discrete information for target recognition, tracking, and situation assessment. In *Proceedings of the SPIE 1699, Signal Processing, Sensor Fusion, and Target Recognition*, 1992.

E. Taghavi, F. Lindsten, L. Svensson, and T. B. Schön. Adaptive stopping for fast particle smoothing. In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, June 1987.

P. Turchin. *Complex Population Dynamics: A Theoretical/Empirical Synthesis*. Princeton University Press, 2003.

J. von Neumann. Various techniques used in connection with random digits. *Journal of Research of the National Bureau of Standards, Applied Mathematics Series*, 12:36–38, 1951.

M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, New York, 2nd edition, 1997.

A. Wills, T. B. Schön, F. Lindsten, and B. Ninness. Estimation of linear systems using a Gibbs sampler. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, July 2012.

Part II

Publications

PhD Dissertations
Division of Automatic Control
Linköping University

M. Millnert: Identification and control of systems subject to abrupt changes. Thesis No. 82, 1982. ISBN 91-7372-542-0.

A. J. M. van Overbeek: On-line structure selection for the identification of multivariable systems. Thesis No. 86, 1982. ISBN 91-7372-586-2.

B. Bengtsson: On some control problems for queues. Thesis No. 87, 1982. ISBN 91-7372-593-5.

S. Ljung: Fast algorithms for integral equations and least squares identification problems. Thesis No. 93, 1983. ISBN 91-7372-641-9.

H. Jonson: A Newton method for solving non-linear optimal control problems with general constraints. Thesis No. 104, 1983. ISBN 91-7372-718-0.

E. Trulsson: Adaptive control based on explicit criterion minimization. Thesis No. 106, 1983. ISBN 91-7372-728-8.

K. Nordström: Uncertainty, robustness and sensitivity reduction in the design of single input control systems. Thesis No. 162, 1987. ISBN 91-7870-170-8.

B. Wahlberg: On the identification and approximation of linear systems. Thesis No. 163, 1987. ISBN 91-7870-175-9.

S. Gunnarsson: Frequency domain aspects of modeling and control in adaptive systems. Thesis No. 194, 1988. ISBN 91-7870-380-8.

A. Isaksson: On system identification in one and two dimensions with signal processing applications. Thesis No. 196, 1988. ISBN 91-7870-383-2.

M. Viberg: Subspace fitting concepts in sensor array processing. Thesis No. 217, 1989. ISBN 91-7870-529-0.

K. Forsman: Constructive commutative algebra in nonlinear control theory. Thesis No. 261, 1991. ISBN 91-7870-827-3.

F. Gustafsson: Estimation of discrete parameters in linear systems. Thesis No. 271, 1992. ISBN 91-7870-876-1.

P. Nagy: Tools for knowledge-based signal processing with applications to system identification. Thesis No. 280, 1992. ISBN 91-7870-962-8.

T. Svensson: Mathematical tools and software for analysis and design of nonlinear control systems. Thesis No. 285, 1992. ISBN 91-7870-989-X.

S. Andersson: On dimension reduction in sensor array signal processing. Thesis No. 290, 1992. ISBN 91-7871-015-4.

H. Hjalmarsson: Aspects on incomplete modeling in system identification. Thesis No. 298, 1993. ISBN 91-7871-070-7.

I. Klein: Automatic synthesis of sequential control schemes. Thesis No. 305, 1993. ISBN 91-7871-090-1.

J.-E. Strömberg: A mode switching modelling philosophy. Thesis No. 353, 1994. ISBN 91-7871-430-3.

K. Wang Chen: Transformation and symbolic calculations in filtering and control. Thesis No. 361, 1994. ISBN 91-7871-467-2.

T. McKelvey: Identification of state-space models from time and frequency data. Thesis No. 380, 1995. ISBN 91-7871-531-8.

J. Sjöberg: Non-linear system identification with neural networks. Thesis No. 381, 1995. ISBN 91-7871-534-2.

R. Germundsson: Symbolic systems – theory, computation and applications. Thesis No. 389, 1995. ISBN 91-7871-578-4.

P. Pucar: Modeling and segmentation using multiple models. Thesis No. 405, 1995. ISBN 91-7871-627-6.

H. Fortell: Algebraic approaches to normal forms and zero dynamics. Thesis No. 407, 1995. ISBN 91-7871-629-2.

A. Helmersson: Methods for robust gain scheduling. Thesis No. 406, 1995. ISBN 91-7871-628-4.

P. Lindskog: Methods, algorithms and tools for system identification based on prior knowledge. Thesis No. 436, 1996. ISBN 91-7871-424-8.

J. Gunnarsson: Symbolic methods and tools for discrete event dynamic systems. Thesis No. 477, 1997. ISBN 91-7871-917-8.

M. Jirstrand: Constructive methods for inequality constraints in control. Thesis No. 527, 1998. ISBN 91-7219-187-2.

U. Forssell: Closed-loop identification: Methods, theory, and applications. Thesis No. 566, 1999. ISBN 91-7219-432-4.

A. Stenman: Model on demand: Algorithms, analysis and applications. Thesis No. 571, 1999. ISBN 91-7219-450-2.

N. Bergman: Recursive Bayesian estimation: Navigation and tracking applications. Thesis No. 579, 1999. ISBN 91-7219-473-1.

K. Edström: Switched bond graphs: Simulation and analysis. Thesis No. 586, 1999. ISBN 91-7219-493-6.

M. Larsson: Behavioral and structural model based approaches to discrete diagnosis. Thesis No. 608, 1999. ISBN 91-7219-615-5.

F. Gunnarsson: Power control in cellular radio systems: Analysis, design and estimation. Thesis No. 623, 2000. ISBN 91-7219-689-0.

V. Einarsson: Model checking methods for mode switching systems. Thesis No. 652, 2000. ISBN 91-7219-836-2.

M. Norrlöf: Iterative learning control: Analysis, design, and experiments. Thesis No. 653, 2000. ISBN 91-7219-837-0.

F. Tjärnström: Variance expressions and model reduction in system identification. Thesis No. 730, 2002. ISBN 91-7373-253-2.

J. Löfberg: Minimax approaches to robust model predictive control. Thesis No. 812, 2003. ISBN 91-7373-622-8.

J. Roll: Local and piecewise affine approaches to system identification. Thesis No. 802, 2003. ISBN 91-7373-608-2.

J. Elbornsson: Analysis, estimation and compensation of mismatch effects in A/D converters. Thesis No. 811, 2003. ISBN 91-7373-621-X.

O. Härkegård: Backstepping and control allocation with applications to flight control. Thesis No. 820, 2003. ISBN 91-7373-647-3.

R. Wallin: Optimization algorithms for system analysis and identification. Thesis No. 919, 2004. ISBN 91-85297-19-4.

D. Lindgren: Projection methods for classification and identification. Thesis No. 915, 2005. ISBN 91-85297-06-2.

R. Karlsson: Particle Filtering for Positioning and Tracking Applications. Thesis No. 924, 2005. ISBN 91-85297-34-8.

J. Jansson: Collision Avoidance Theory with Applications to Automotive Collision Mitigation. Thesis No. 950, 2005. ISBN 91-85299-45-6.

E. Geijer Lundin: Uplink Load in CDMA Cellular Radio Systems. Thesis No. 977, 2005. ISBN 91-85457-49-3.

M. Enqvist: Linear Models of Nonlinear Systems. Thesis No. 985, 2005. ISBN 91-85457-64-7.

T. B. Schön: Estimation of Nonlinear Dynamic Systems — Theory and Applications. Thesis No. 998, 2006. ISBN 91-85497-03-7.

I. Lind: Regressor and Structure Selection — Uses of ANOVA in System Identification. Thesis No. 1012, 2006. ISBN 91-85523-98-4.

J. Gillberg: Frequency Domain Identification of Continuous-Time Systems Reconstruction and Robustness. Thesis No. 1031, 2006. ISBN 91-85523-34-8.

M. Gerdin: Identification and Estimation for Models Described by Differential-Algebraic Equations. Thesis No. 1046, 2006. ISBN 91-85643-87-4.

C. Grönwall: Ground Object Recognition using Laser Radar Data – Geometric Fitting, Performance Analysis, and Applications. Thesis No. 1055, 2006. ISBN 91-85643-53-X.

A. Eidehall: Tracking and threat assessment for automotive collision avoidance. Thesis No. 1066, 2007. ISBN 91-85643-10-6.

F. Eng: Non-Uniform Sampling in Statistical Signal Processing. Thesis No. 1082, 2007. ISBN 978-91-85715-49-7.

E. Wernholt: Multivariable Frequency-Domain Identification of Industrial Robots. Thesis No. 1138, 2007. ISBN 978-91-85895-72-4.

D. Axehill: Integer Quadratic Programming for Control and Communication. Thesis No. 1158, 2008. ISBN 978-91-85523-03-0.

G. Hendeby: Performance and Implementation Aspects of Nonlinear Filtering. Thesis No. 1161, 2008. ISBN 978-91-7393-979-9.

J. Sjöberg: Optimal Control and Model Reduction of Nonlinear DAE Models. Thesis No. 1166, 2008. ISBN 978-91-7393-964-5.

D. Törnqvist: Estimation and Detection with Applications to Navigation. Thesis No. 1216, 2008. ISBN 978-91-7393-785-6.

P-J. Nordlund: Efficient Estimation and Detection Methods for Airborne Applications. Thesis No. 1231, 2008. ISBN 978-91-7393-720-7.

H. Tidefelt: Differential-algebraic equations and matrix-valued singular perturbation. Thesis No. 1292, 2009. ISBN 978-91-7393-479-4.

H. Ohlsson: Regularization for Sparseness and Smoothness — Applications in System Identification and Signal Processing. Thesis No. 1351, 2010. ISBN 978-91-7393-287-5.

S. Moberg: Modeling and Control of Flexible Manipulators. Thesis No. 1349, 2010. ISBN 978-91-7393-289-9.

J. Wallén: Estimation-based iterative learning control. Thesis No. 1358, 2011. ISBN 978-91-7393-255-4.

J. Hol: Sensor Fusion and Calibration of Inertial Sensors, Vision, Ultra-Wideband and GPS. Thesis No. 1368, 2011. ISBN 978-91-7393-197-7.

D. Ankelhed: On the Design of Low Order H-infinity Controllers. Thesis No. 1371, 2011. ISBN 978-91-7393-157-1.

C. Lundquist: Sensor Fusion for Automotive Applications. Thesis No. 1409, 2011. ISBN 978-91-7393-023-9.

P. Skoglar: Tracking and Planning for Surveillance Applications. Thesis No. 1432, 2012. ISBN 978-91-7519-941-2.

K. Granström: Extended target tracking using PHD filters. Thesis No. 1476, 2012. ISBN 978-91-7519-796-8.

C. Lyzell: Structural Reformulations in System Identification. Thesis No. 1475, 2012. ISBN 978-91-7519-800-2.

J. Callmer: Autonomous Localization in Unknown Environments. Thesis No. 1520, 2013. ISBN 978-91-7519-620-6.

D. Petersson: A Nonlinear Optimization Approach to H2-Optimal Modeling and Control. Thesis No. 1528, 2013. ISBN 978-91-7519-567-4.

Z. Sjanic: Navigation and Mapping for Aerial Vehicles Based on Inertial and Imaging Sensors. Thesis No. 1533, 2013. ISBN 978-91-7519-553-7.