UNIVERSITY COLLEGE LONDON

Approximate Inference for State-Space Models

by

Matthew Charles Higgs

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of engineering
Computer science

December 2011

Declaration of Authorship

I, Matthew Charles Higgs, declare that this thesis titled, 'Approximate Inference for State-Space Models' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Abstract

This thesis is concerned with state estimation in partially observed diffusion processes with discrete time observations. This problem can be solved exactly in a Bayesian framework, up to a set of generally intractable stochastic partial differential equations. Numerous approximate inference methods exist to tackle the problem in a practical way. This thesis introduces a novel deterministic approach that can capture non normal properties of the exact Bayesian solution.

The variational approach to approximate inference has a natural formulation for partially observed diffusion processes. In the variational framework, the exact Bayesian solution is the optimal variational solution and, as a consequence, all variational approximations have a universal ordering in terms of optimality. The new approach generalises the current variational Gaussian process approximation algorithm, and therefore provides a method for obtaining super optimal algorithms in relation to the current state-of-the-art variational methods.

Every diffusion process is composed of a drift component and a diffusion component. To obtain a variational formulation, the diffusion component must be fixed. Subsequently, the exact Bayesian solution and all variational approximations are characterised by their drift component. To use a particular class of drift, the variational formulation requires a closed form for the family of marginal densities generated by diffusion processes with drift components from the aforementioned class. This requirement in general cannot be met. In this thesis, it is shown how this coupling can be weakened, allowing for more flexible relations between the variational drift and the variational approximations of the marginal densities of the true posterior process. Based on this revelation, a selection of novel variational drift components are proposed.

Acknowledgements

First and foremost, thanks to my supervisor John Shawe-Taylor for giving me unprecedented freedom throughout my PhD. For being insightful when we met and for providing me with opportunities that many PhD students long for. Thanks to Steffen Grunewalder and Guy Lever for their constant wayward swing to mathematics. For being pro-active with work and for always being attracted to ambitious problems. Thanks to the CSML group: Louis Dorard, David Hardoon, Cedric Archambeau, Zakira Hussain, Tom Diethe, and Tom Furmston. Also thanks to Manfred Opper and Francois Laviolette, and to Michail Vrettas for some very helpful code.

Contents

Dedicated to my parents.

Chapter 1

Introduction

1.1 Purpose and contributions of the thesis

The purpose of this thesis is to improve upon existing variational approximate inference methods for state-estimation in partially observed diffusion processes with discrete time observations. The emphasis is on continuous-discrete state space models with nonlinear drift functions and measurement models, in which the state estimation problem cannot be solved using classical linear methods. The mathematical approach is probabilistic and the state estimation problem is formulated in terms of Bayesian inference. All algorithms are considered in the context of machine learning. This requires a much wider scope than is usually considered for this type of problem, subsuming the more specialised filtering and smoothing algorithms. This is done because many nonlinear smoothing algorithms tackle the state estimation problem by simply adapting procedures that work in the linear case. Effective state estimation in nonlinear models requires a more flexible algorithmic framework, and while many of the ideas in machine learning were seeded in the filtering and smoothing community, progress in other fields of machine learning has created a rich toolbox of approximate inference methods applicable to the problem studied here.

The posterior in any nonlinear model is non normal. Despite this, many deterministic approximate inference methods are implemented as normal approximations. Machine learning algorithms such as expectation propagation (Minka, 2001), assumed density filtering (Opper, 1998), assume density smoothing (Särkkä & Hartikainen, 2010a), and variational free energy methods (Archambeau & Opper, 2011) are all applied to temporal state estimation problems using normal densities as the approximating class. This generally false assumption is then exaggerated further by introducing normal based integral approximations such as statistical linearisation (Roberts & Spanos, 2003), unscented transforms (Julier & Uhlmann, 2004), quadrature methods (Särkkä & Hartikainen, 2010b) and cubature methods (Arasaratnam & Haykin, 2009). While there are many benefits to choosing normal approximations, this does not escape the fact that all nonlinear partially observed diffusion processes will have posteriors with marginal densities better approximated by more general classes of probability densities. While arbitrary approximating densities can be fitted using the samples in stochastic approximate inference methods, the purpose of this thesis is to give some indication of how to do this in the deterministic setting.

At the core of the thesis is the understanding that the true Bayesian solution should be used to guide any approximation. The Fokker-Planck (Risken, 1996) or Kolmogorov forward equation (Kolmogorov, 1931) is the fundamental stochastic partial differential equation that couples the drift of a diffusion process to its corresponding measure. Conversion of a continuous time system to a numerically tractable discrete time system requires the Fokker-Planck equation to produce a sequence of transition probabilities that characterise the propagation of beliefs between disjoint times on the grid. In general this equation is intractable, and learning algorithms must approximate the transition densities and perform inference on a simplified version of the prior. Certain stochastic methods attempt to reduce the impact of approximating the transition density by using high order numerical approximations (Murray & Storkey, 2011, Restrepo, 2008). While some deterministic methods derive continuous time formulations by taking the time step of the discretisation to zero (Särkkä, 2007, Särkkä & Sarmavouri, 2011). It is only the variational methods studied here that have the ability to integrate the Fokker-Planck equation explicitly into the formulation of solutions.

1.1.1 Contributions

The contributions of the thesis are summarised as follows:

- **Generalised variational smoothing:**

  The variational Gaussian process approximation for partially observed diffusion processes with discrete time observations is extended to any family of densities parameterised by a finite set of moments. This allows for variational approximations to capture higher order moments of the true posterior process.

- **Projected variational smoothing:**

  The strict constraints of generalised variational smoothing are relaxed through the introduction of a projected version of the Fokker-Planck equation with a novel proof in terms of the continuous time limit of an assumed density filter. This allows for more flexible types of variational drift to be introduced.

- **Variational drifts with additive control:**

  Based on the new framework, a novel form of variational drift is proposed that plays upon the optimal control and estimation duality. It is shown how the variational GP approximation can be interpreted in this framework. The hope is to strengthen the link between approximate estimation and approximate control.

- **Variational gradient systems:**

Based on the new framework, a novel form of variational drift is proposed that focuses on the equilibrium points of the Fokker-Planck equation. It is shown how gradient systems can be used to build well behaved vector fields that guide the marginal densities of the variational approximation. The new type of drift encompasses a large number of models encountered in applications.

- **Variational skew-GP smoothing:**

  A method is proposed for capturing skewness in filtering and variational smoothing. It is shown how the class of skew-normal distributions can be adapted to the task of approximate inference. Many of the auxiliary and prerequisite steps required for variational skew-GP smoothing are dealt with.

- **Variational GP learning:**

  The variational GP algorithm is considered in the context of other temporal GP learning machine. The formal relations between variational GP smoothing and GP regression, kernel regression, and assumed Gaussian smoothing are identified. Improvements to the current gradient method used to implement the variational GP method are proposed. These improvements utilise the Hessian of the free-energy, which has yet to be used in variational GP smoothing.

Work not included in this thesis comprises of active and dormant collaborations in optimal bandit control (Grunewalder et al., 2010), statistical learning theory for variational approximations (Archambeau et al., 2009), a first-author conference publication on PAC-Bayesian theory for density estimation (Higgs & Shawe-Taylor, 2010), and a first-author workshop publication on multiple kernel learning for system identification (Higgs & Shawe-Taylor., 2010). Rather than try and combine these disparate topics, the decision was made to write a complete, contained and (hopefully) enjoyable to read thesis. Current and future paper proposals based on the contents of this thesis are discussed at the close.

1.1.2 Scope of thesis

The thesis approaches inference in nonlinear state-space models from a machine learning perspective. While alternative methodologies for inference in state-space models are considered, all of the algorithms in the thesis fall into a specific class. To show this, figure 1.1 illustrates a highly simplified map of machine learning algorithms. The two main variables (in orange) are the type of the model and the unknown variables in the model to be learned. Hierarchal dynamic models (HDMs) (Friston, 2008a) consists of hierarchies of static and dynamic interacting layers of probabilistic models up to any depth required. Here only the shallow dynamic subclass of state-space models are considered. The variables of the Bayesian model are split into states and parameters. This distinction can blur somewhat, especially when parameters are incorporated the set of unknowns as additional states and learned in tandem with the other latent

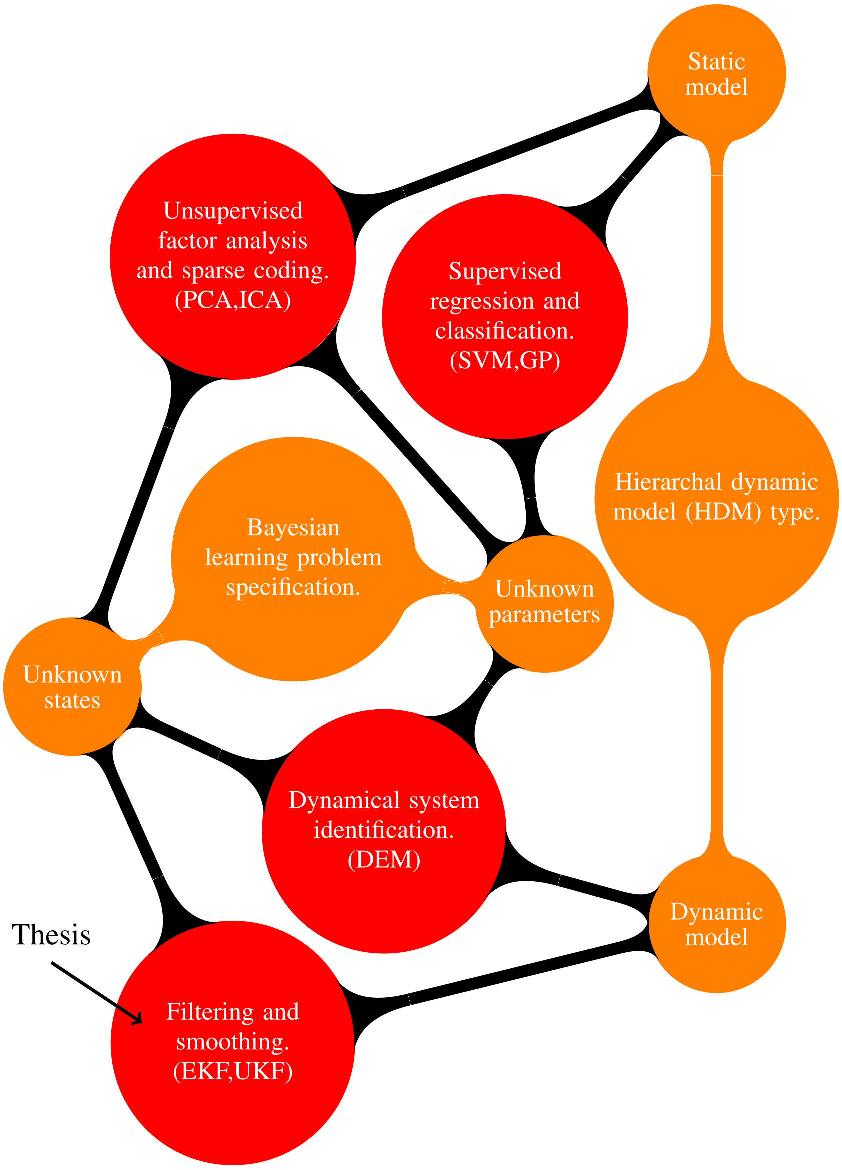

Static model
Unsupervised factor analysis and sparse coding. (PCA,ICA)
Supervised regression and classification. (SVM,GP)
Hierarchal dynamic model (HDM) type.
Bayesian learning problem specification.
Unknown parameters
Unknown states
Dynamical system identification. (DEM)
Dynamic model
Thesis
Filtering and smoothing. (EKF,UKF)

FIGURE 1.1: Simplified map of machine learning algorithms. Different combinations of the two main variables (in orange) lead to different machine learning problems (in red). Abbreviations are as follows: Principle component analysis (PCA), Independent component analysis (ICA), Support-vector machine (SVM), Gaussian processes (GP), Dynamic Expectation Maximisation (DEM), Extended Kalman filter (EKF), Unscented Kalman filter (UKF).

variables. For now, states can be thought of as variables that undergo mappings or evolve over time. Parameters can be thought of as properties of these mappings. How the problem specifics (type of model and set of unknown variables) are combined, leads to a highly simplified taxonomy of machine learning algorithms. When states and parameters are unknown, the problem is considered unsupervised. In a static setting, examples of unsupervised learning algorithms are dimensionality reduction, cluster analysis, and independent component analysis (Ghahramani, 2004). In a dynamic setting unsupervised algorithms can be collectively considered as system identification algorithms, and require dynamic states and parameters to be learned in tandem. Static models with unknown parameters lead to traditional supervised and neural-network algorithms that focus on regression and classification problems. Dynamic models with unknown states lead to the traditional filtering and smoothing algorithms. Thus the thesis considers algorithms that would be at home in the lower left region of the figure. The learning problem is how to infer hidden or latent states, given a state-space model with known parameters. This leads to smoothing algorithms, most of which are easily extended to parameter estimation and system identification. The algorithm abbreviations in the brackets in figure 1.1 are common examples of algorithms used to tackle the statistical learning problems on high. They are not exclusive and the list is not exhaustive. Indeed, it will be discussed later on in the thesis how many of the Gaussian Process regression algorithms can be (and frequently have been) applied to temporal smoothing problems.

1.1.3 Outline of thesis

The thesis begins with a short discussion of the applications of optimal smoothing and its origins in the filtering and smoothing community. This is followed by a discussion of general approximate inference and its specialisation to the smoothing problem. These topics are covered in sections 1.2 and 1.3, respectively.

In chapter 2 the model is specified in complete detail, collating the required components from Bayesian inference and stochastic processes and introducing notation. In subsection 2.5 an overview of machine learning approximate inference methods is given, with each approach discussed in the context of state space models. The purpose is to expose the current state-of-art algorithms in implementable form, and to expose the place of variational methods in this mix. The exposition also gives an inclination of what it means to be a continuous time algorithm with regards to the temporal state-estimation problem.

In chapter 3 the variational GP smoothing framework of Archambeau & Opper (2011) is generalised to any family of densities parameterised by a finite set of moments. First the optimal solution in a variational formulation is shown to be equivalent to the exact Bayesian solution. It is then shown how suboptimal variational approximations can be constructed by restricting the variational drift and corresponding marginal densities to a tractable class. This allows the drift-marginal relations to be captured through simple moment equations, which leads to a general variational smoothing algorithm that subsumes the GP algorithm of Archambeau & Opper (2011). The generalised framework requires an explicit form for the drift of a diffusion process and the corresponding set of marginal densities, which are generally intractable. Section 3.4 introduces a projection method that enables intractable marginal densities to be projected onto a tractable class. The traditional formulation of this projection requires some in-depth differential geometry, and in subsection 3.5 a novel proof is given that combines the proof of the Fokker-Planck equation with the derivation of the discrete time assumed density filter (Opper, 1998).

Subsection 3.6 discusses how the projection method can be incorporated into the general variational smoothing framework to enable the use of more relaxed relations between the variational drift and the variational approximation of the marginal densities of the true posterior process.

In chapter 4 the generalised framework and its projective relaxation is used to propose some novel drift functions for use in the variational smoothing algorithm. Subsection 4.1 outlines the basics of the stochastic optimal control problem and considers its relation to the variational formulation of exact inference. Subsection 4.1.2 discusses the connection between the variational smoothing algorithm and variational optimal controls and the variational GP algorithm is interpreted in the language of optimal control. Subsection (4.2) discusses gradient systems and their time-varying extensions and how they can be used in the variational framework. Subsection 4.3 considers the practical aspects of capturing and tracking higher order moments. Subsection 4.3.1 reviews the skew-normal density and how it could be used in filtering and variational smoothing.

In chapter 5, attention moves back to the optimal Gaussian variational smoothing solution. The first part of the chapter, section 5.3.1 places the optimal Gaussian variational solution in the context of other temporal Gaussian process algorithms, highlighting the connections between the stationary kernels common to spatial methods in machine learning, related reproducing kernel type methods and recent assumed density smoothing methods. In section 5.4, important implementation details are discussed. A review of the conjugate gradient methods used for learning the variational parameters of the optimal Gaussian smoothing algorithm is given, and an analysis of the Hessian of the free-energy of the Gaussian approximation suggests improvements that can be made. Dealing with intractable expectations is discussed, and how these can be extended higher order moments and more general marginal densities.

The final chapter of the thesis reviews the key points of the thesis and the many unsolved problems and directions of future work. Section 6.1 discusses machine learning algorithms with continuous time formulations and some alternative variational approaches from other fields such as neuroscience and geophysics. Paper proposals and future work are discussed and final thoughts are given.

1.2 Optimal smoothing

Optimal smoothing is interpreted in probability theory as the conditioning of time varying random states on a set of noisy observations. Importantly, the one time marginal densities that follow from optimal smoothing are functions of all the data available in a fixed time window. This is in contrast to optimal filtering, which considers one time marginal densities condition only on data from earlier times. Some applications of optimal smoothing are given below.

1.2.1 Applications of optimal smoothing

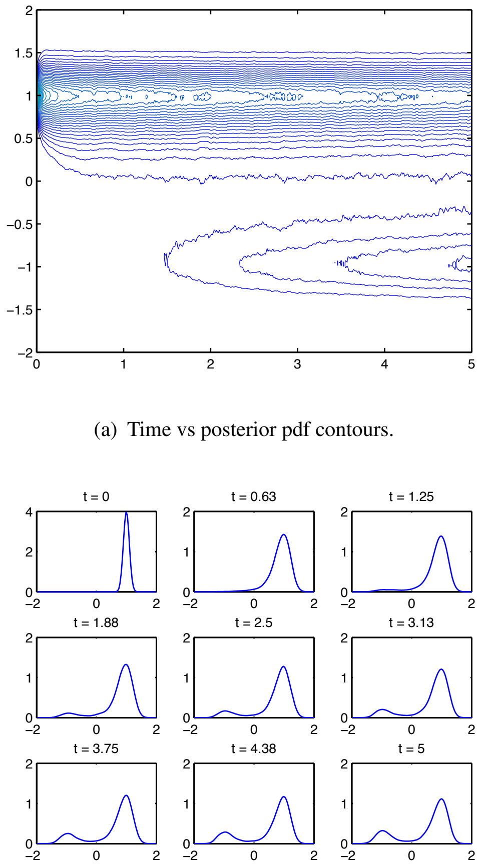
- *Functional Magnetic Resonance Imaging* measures hemodynamic responses, which represent changes in blood flow and blood oxygenation that follow neuronal activation. The interaction between neuronal activity and observations rests on a well known electro/biophysical process (Riera et al., 2004). Smoothing methods (Havlicek et al., 2011, Murray & Storkey, 2011) allow the model to be inverted to generate estimates on the hidden states of neuronal activity.

- *Weather prediction* involves the numerical modelling of complex ocean and climate dynamics. While forecasting is an obvious problem in this setting, hindcasting and smoothing methods (Eyink et al., 2004, Miller et al., 1999) are used to solve the inverse problem for complex nonlinear quasi-geostrophic systems.

- *Stochastic optimal control* has applications in areas of biology (Joshi, 2002) and robotics (Theodorou et al., 2010), among many others. Particular stochastic optimal control problems can be phrased as continuous time optimal smoothing problems, allowing for a large amount of approximate inference smoothing algorithms to be transferred to problems of optimal control.

1.2.2 Non normality in general models

The main property of the above applications, is the nonlinearity of the underlying dynamics. Classical nonlinear approximation methods, such as the extended Kalman filter, have been shown to fail on such tasks (Miller et al., 1999). When observations are few and far between the nonlinear dynamics of the prior quickly leads to conditional densities that are far from Gaussian. Shown in figure 1.2 (a) is a plot of the time-evolution of the posterior of the double-well system with a noise-free observation at time zero. The double-well system has become the benchmark example in nonlinear data assimilation (Archambeau et al., 2007a, Eyink et al., 2004, Miller et al., 1999). The time-slice plots in 1.2 (b) give an idea of how the shape of the posterior changes over time as it converges towards the bimodal equilibrium distribution of the double-well system. For times directly after the observation the posterior is essentially Gaussian, but this quickly changes as the posterior begins to take on a skewed and then bimodal form as probability mass flows into the adjacent well. If accurate representation of the posterior is the primary goal, then Gaussian approximations are only sufficient for systems with high frequency observations and low observation noise. For more general systems, more general approximating classes are required. While arbitrary moments can be captured using the samples in stochastic methods, the main foci of this thesis is how to do this in the deterministic setting.



(a) Time vs posterior pdf contours.

(b) One-time-slice marginal densities.

FIGURE 1.2: Plots of time-evolution of posterior density (a), and one-time-slice marginal densities at a selection of times (b). The posterior is built from the *double-well* steady-state prior and a noise-free observation at $x(0) = 1$.

1.3 Practical learning methods

The action of assimilating data into prior beliefs is most naturally formulated in the language of Bayesian inference. Recent efforts in Bayesian analysis have been focused on algorithmic implementation. The need to construct a practical learning algorithm affects the criteria of what we consider optimal, and while approximate inference is often necessary in learning due to intractable integrals, it is also often desired for improved computational complexity. In nonlinear models, the only way to solve the Bayesian inference problem is by approximation, and therefore all progress over the past few decades has been in providing efficient alternative methods to exact inference. These methods are designed to both capture the data and to preserve the relevant properties of the prior. For the particular case of state-space models, exact solutions for the state posterior were finalised in the 1960-70's (Leondes et al., 1970, Striebel, 1965), but their intractability in the general case was known even then.

1.3.1 Approximate inference methods

Approximate inference can be naively divided into stochastic and deterministic methods (see figure 1.3). Stochastic methods utilise the generative model to draw samples from the intractable

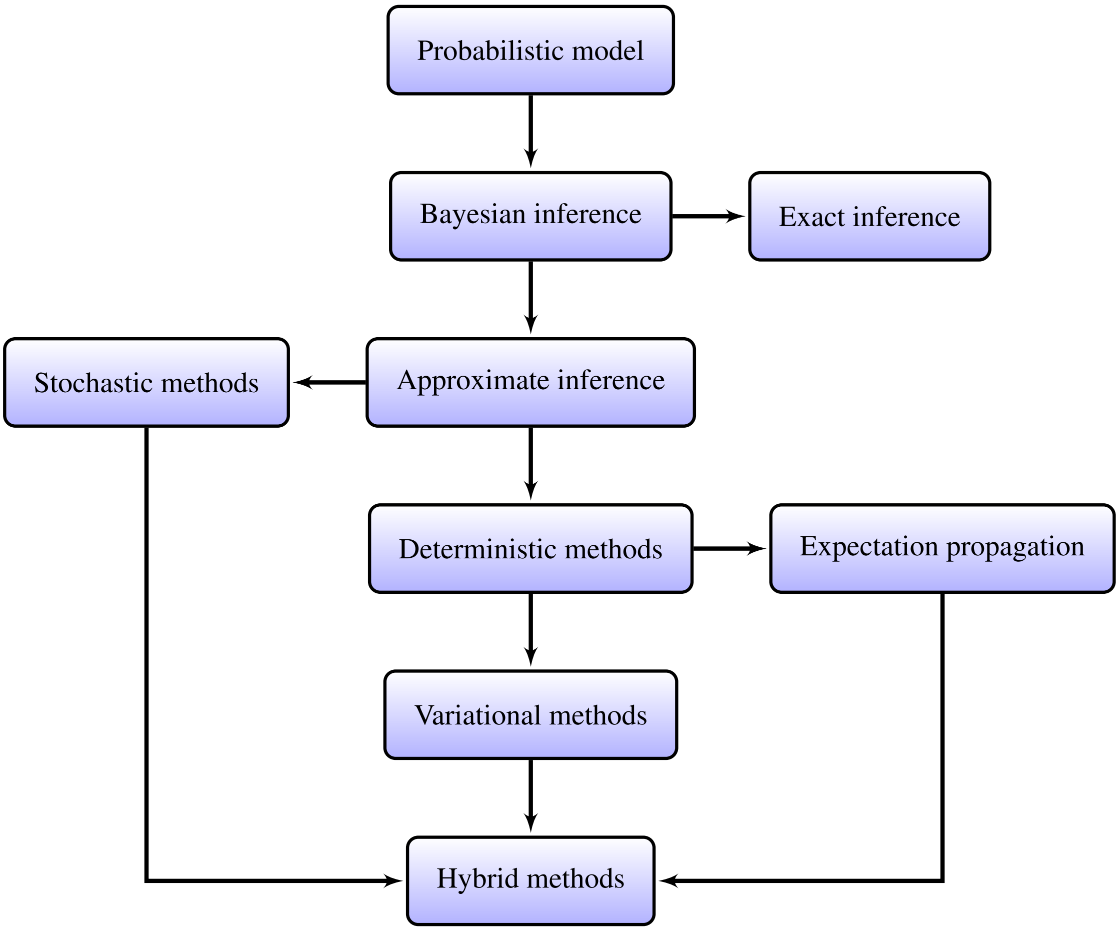


FIGURE 1.3: Simplified hierarchy for practical Bayesian inference.

posterior and their validity relies upon statistical guarantees for the convergence of empirical estimations. In contrast, deterministic methods look to approximate the intractable posterior with a distribution from a simpler more tractable model. This leads naturally to a well defined criteria of optimality by considering a dissimilarity measure between the posterior and the approximation. While the primary focus of the thesis is on deterministic methods, namely how can the state-of-the-art be improved in accuracy and efficiency, stochastic or Monte Carlo methods still feature strongly. The natural synergy of Monte Carlo methods with generative models likens them to the methods used for generating synthetic data. They are an important part of hybrid algorithms, allowing many restrictive assumptions to be relaxed. Additionally, with enough samples Monte Carlo methods can be considered a computationally expensive benchmark against which to compare competing deterministic methods. In the discrete-time setting, various deterministic approximation methods have been developed to deal with the intractable integrals that follow from a recursive Bayesian analysis in discrete time state-space models. All of these recursive discrete time methods are essentially nonlinear extensions of the classical Kalman filter (Kalman, 1960). They utilises the probabilistic structure of the model, dealing with nonlinear mappings by transferring only the most relevant statistics of the current belief. They are sequential moment matching methods and at the turn of the millennia seeded the more general framework known as expectation propagation (Heskes & Zoeter, 2002, Minka, 1999, 2001), which quickly became popular in machine learning due to its simplicity and generality. An alternative framework for deterministic approximate inference, with its origins in statistical physics, considers the classical expectation-maximisation algorithm of Dempster et al. (1977) as a computationally tractable way to minimise a free-energy functional bounding the relative entropy between the posterior and an approximation (Neal & Hinton, 1998). The framework allows hidden states and unknown parameters to be learned in tandem in a principled fashion, and many learning schemes are subsumed by the variational free-energy approach. The appeal of formulating approximate inference as an optimisation problem has also made variational methods hugely popular throughout machine learning and its related fields (Jaakkola, 2000, Jordan et al., 1999a).

1.3.2 Learning in continuous time

Many real world systems are naturally modelled in continuous time. Unfortunately the default solution in many areas of machine learning is to apply a first order approximation to the prior model. This enables well established discrete-time algorithms to be applied "out-of-the-box", but can lead to implicit errors, possible computational costs, and a lack of descriptive power with regards to jumps and transitions. While the current world of digital computers forces us to discretise time at some point in the computational implementation of any continuous-time inference procedure, the guiding principle in numerical analysis and its applications is to avoid discretisation until the last possible moment (Stuart, 2010). This is reflected in Bayesian inference by the desire to use higher order discretisation schemes for continuous-time models. In spite of this, the use of continuous-time models in Bayesian inference is severely underdeveloped. In sampling-based approximate inference methods, linear discretisation of the prior is often the first step, and only recently have "continuous-time" sampling-based methods matured into useable algorithms (Murray & Storkey, 2011). These algorithms allow particles to be propagated through time using faster and more accurate numerical stochastic-integration schemes with variable time scales. In deterministic approximate inference, continuous-time solutions are desirable for similar reasons. The moments of posterior approximations inherit deterministic differential properties of the prior. This allows posterior models to be learned, and posterior solutions to be generated at will using the huge resource of deterministic numerical methods. Due to the classical path-integral formulation in physics (Feynman & Hibbs, 1965), variational inference methods have a natural formulation in continuous-time and these continuous-time free-energy formulations have become popular in machine learning (Archambeau et al., 2007a, Archambeau & Opper, 2011, Archambeau et al., 2007b), neuroscience (Friston, 2008a,b, Friston et al., 2008) and the more general (but geophysically motivated) field of data assimilation (Eyink, 2000, Eyink et al., 2004). The continuous-time analogue for moment matching methods is not so obvious, and is considered at the close in section 6.1.

Any stochastic-process related investigation aimed at researchers and practitioners in machine learning has a chamber of measure theory and infinite dimensional analysis bubbling under the surface. A fully rigorous exposition is generally avoided for shear readability and the approach here follows a balance between rigour and readability seen in many common physics, neuroscience and machine learning papers. For example there is no notational distinction between random variables and their realisations (commonly denoted by upper case X and lower case x, respectively). At times though, certain rigour is required. The use of "white-noise" (common in physics) is avoided, and stochastic differential equations (SDE)'s are expressed in Itô or Stratonovich form (Øksendal, 2003). This is important, not only because it is the correct way to express an SDE, but because many of the results in the continuous-time setting are sensitive to how the equations of motion are defined. These mathematical technicalities are as much a part of the model as the parts describing the high level behaviours that fit our intuitions.

Chapter 2

Background

2.1 Introduction

This section covers the basics of approximate Bayesian inference methods in machine learning, with directed attention to nonlinear state-space models. The fundamental inference problem is outlined in section 2.2 with a short discussion on the technicalities of dealing with infinite dimensional latent variables. Section (2.2.2) discusses some fundamental mathematical properties of the state-space model and the underdetermined nature of solutions. The ideas are illustrated by example using Brownian motion, the driving force in all the models considered in this thesis. Section 2.3 specifies the model in detail. All of the required notation is introduced and many auxiliary results underlying the properties of the model are discussed. General results for stochastic processes are taken from Øksendal (2003) and Kuo (2006). Results more focused on the Fokker-planck equation are taken from Risken (1996) and Kuo (2006). Section (2.3.4) introduces the required tools to describe general diffusion models. The section explains how partially observed diffusion processes can be solved exactly in a Bayesian framework, up to a set of (generally intractable) stochastic partial-differential equations. Section 2.5 discusses approximate inference methods in machine learning. The section gives an overview of recursive Monte Carlo, expectation propagation, assumed density methods, Hybrid Monte Carlo and variational methods specialised to inference in state-space models. The chapter finishes with a short discussion of finite dimensional exponential families.

2.2 Latent model inversions

Consider a simple generative model $p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})$, where $\mathbf{X}$ is hidden and $\mathbf{Y}$ is observed and both are Lebesgue measurable. There are various reasons why one may want to use hidden variables: maybe to compress the description of the data by methods such as dimensionality reduction, or to allow complex distributions to be expressed in terms of more tractable joint distributions over extended parameter spaces (Bishop, 1999). The role of latent variable models in this thesis is to encode complex prior information, and to enable the assimilation of data into these prior beliefs. Under Bayes' rule the generative model can be inverted to give a posterior distribution

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})} \tag{2.1}$$

over the hidden variables $\mathbf{X}$ given the observable $\mathbf{Y}$. While equation (2.1) has no obvious temporal attributes, it is the cornerstone of learning with time-series data. Under suitable assumptions, inference over hidden states can be restricted to isolated time-windows or individual instances of time. Earlier inferences are encapsulated in time dependent priors and equation (2.1) enables data points to be assimilated into current beliefs. To do this, the probabilistic dependencies need to be given concrete forms. A common representation of the conditional probability $p(\mathbf{Y}|\mathbf{X})$ is through an observation model

$$\mathbf{Y} = \mathfrak{H}(\mathbf{X}) + \boldsymbol{\eta} \tag{2.2}$$

where $\mathfrak{H} : \mathbf{X} \mapsto \mathbf{Y}$ is a nonlinear observation operator and $\boldsymbol{\eta}$ is a zero-mean random variable with probability density ϕ. This type of model forms the core of Bayesian inverse problems (Stuart, 2010). While a rigorous representation of the system would include '$\boldsymbol{\eta}$' in the generative model, such auxiliary variables are omitted to ease notation. Now assume $\mathbf{X}$ inhabits a general measurable space, and let P_{prior} and P_{post} denote probability measures relating to the laws of $p(\mathbf{X})$ and $p(\mathbf{X}|\mathbf{Y})$ respectively. The prior measure, posterior measure, and likelihood function $\phi\big(\mathbf{Y} - \mathfrak{H}(\mathbf{X})\big)$ are all related according to Bayes' rule through the Radon-Nikodym derivative

$$\frac{dP_{post}}{dP_{prior}}(\mathbf{X}) \propto \phi\big(\mathbf{Y} - \mathfrak{H}(\mathbf{X})\big). \tag{2.3}$$

While equations (2.1) and (2.3) are essentially the same, the use of abstract measures in (2.3) is key. Many real world systems are described by models with hidden states indexed by continuous-valued variables. Representation of beliefs and learning in these models requires the delicate handling of probability measures on infinite dimensional spaces. While such models bring unrivalled descriptive power, they come with many technical issues and computational complexities. The intuitive appeal of sample paths evolving through state-space can be deceiving of the tough work required to perform practical inference in continuous-time.

2.2.1 Learning from time-series

In this thesis the observable $\mathbf{Y}$ is given by an ordered set $\{\mathbf{y}_i \in \mathbb{R}^{d_\mathbf{y}}, t_i \in \mathcal{T}\}$ for some ascending finite subset $\mathcal{T} \subset [0, T]$, referred to as a time-series. The observation operator $\mathfrak{H}$ is given by $\mathfrak{H}(\mathbf{x}) = \{\mathbf{h}(\mathbf{x}(t_i), t_i), t_i \in \mathcal{T}\}$ for some map $\mathbf{h} : \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}_+ \to \mathbb{R}^{d_\mathbf{y}}$. In this description the indices of times in $\mathcal{T}$ are used to match[1] the outputs of $\mathfrak{H}$ to the appropriate observations. The formulation allows for missing observations and is referred to as continuous-discrete, corresponding to the continuous-time nature of the hidden states and the discrete-time nature of the

[1] There is a more formal way of defining $\mathfrak{H}$ in both these settings using a projection of the hidden state onto the observation times. This requires additional technical detail without much pedagogical gain.

observations. It is also possible for $\mathbf{Y}$ to be infinite dimensional, e.g. defined by a continuous-time map $\mathbf{y}(t)$. This is very common in traditional filtering problems (Øksendal, 2003), but requires some more involved mathematics and is not relevant to the setting here.

The type of learning studied here is distinct from the more common category of supervised learning. These differences can be easily seen by considering the observation model in (2.2). This is in effect the simplest form of model consisting of "causes" $\mathbf{X}$, "effects" $\mathbf{Y}$, and an operator $\mathfrak{H}$ between the two. In the batch setting of supervised learning, $\mathbf{X}$ and $\mathbf{Y}$ correspond to a set of causes $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and effects $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ and the observation operator $\mathfrak{H}$ is given by $\mathfrak{H}(\mathbf{X}) = (\mathbf{h}(\mathbf{x}_1), \ldots, \mathbf{h}(\mathbf{x}_n))$ for some map $\mathbf{h} : \mathbf{x} \mapsto \mathbf{y}$. (traditional) supervised learning assumes the n examples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ are drawn from a joint distribution $p(\mathbf{X}, \mathbf{Y})$, and the task is to utilise the information in the training sample, and a prior on $\mathbf{h}$ or its parameters, to learn a posterior over the mapping $\mathbf{h}$. An obvious example of supervised learning that fits this model is regression. In contrast to this, in the latent-model-inversion problem considered here only the set $\{\mathbf{y}_i\}_{i=1}^{n}$ is observed. Samples are not assumed i.i.d. and the unobserved causes $\mathbf{X}$ can occupy any space the model requires. The approach borders on unsupervised learning, depending on how much of the model is fixed. For the case when $\mathbf{h}$ is assumed known, as is common in many filtering and tracking applications, learning methods try to estimate the hidden states $\mathbf{X}$. When $\mathbf{h}$ is unknown, learning is a dual estimation problem over states $\mathbf{X}$ and the mapping $\mathbf{h}$. This leads to an important distinction between the identification of a model's parameters and estimation of its hidden states.

2.2.2 Data assimilation

The type of learning considered in the atmospheric and oceanographic sciences, geoscience and neuroscience, deals with assimilating data into complex models. The name "data assimilation" emphasises the model's dominance over the data, and a key problem when dealing with such models is the underdetermined nature of solutions. This is due to the sometimes behemothic difference between the dimensions of the hidden and observable variables. This problem is nowhere more apparent than in continuous-discrete state space models, where infinite dimensional hidden variables are pinned down by only a few observations. The problem of underdetermined solutions plagues many complex inverse problems, but its properties are easily captured in a simple example.

2.2.2.1 Simple Brownian motion

Let $\{x(t)\}_{t \in [0,T]}$ denote a $\mathbb{R}$-valued stochastic process evolving according to $dx_t = \sqrt{\sigma} dw_t$, where dw_t is the standard Wiener process in $\mathbb{R}$ (i.e. Brownian motion) scaled by $\sqrt{\sigma} > 0$. Assume the initial condition $x(0) = 0$. This simplified model represents a vanilla version of the problems studied in this thesis. It allows us to consider some fundamental mathematical properties of the model without being clouded by the more sophisticated nonlinear inference

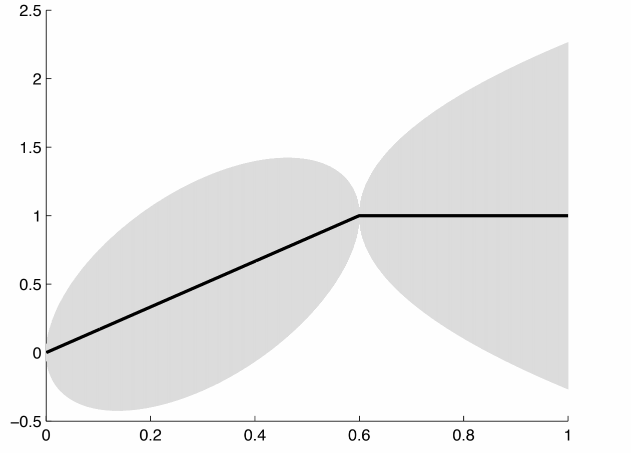


FIGURE 2.1: Plot of Gaussian posterior approximation for brownian motion with initial condition $x(0) = 0$ and a noise-free observation $x(0.6) = 1$. x-axis represents time $t \in [0, 1]$. The mean $\mu^+(t)$ in (2.6) is plotted in black, the confidence region $\{y|(y - \mu^+(t))^2 \leq \kappa^+(t, t)\}$ is shaded in grey.

machinery. For this example there is no notion of drift, the prior model encodes the belief that $x(t)$ moves under Brownian motion with covariance function $k(\tau, t) = \sigma \min\{\tau, t\}$ and mean $\nu = 0$. Consider an observation model $y = x(s)+\eta$, for some $s \in [0, T]$ and $\eta \sim \mathcal{N}(0, r)$. Thus we are considering just one observation in the interval $[0, T]$. Letting $\mu(t)$ and $\kappa(s, t)$ denote the posterior mean and covariance function, respectively, using theorem A.2 (appendix A.3.1) it holds that

$$\mu(t) := \frac{yk(s,t)}{k(s,s)+r}, \qquad \kappa(s,t) := k(t,\tau) - \frac{k(s,t)k(\tau,s)}{k(s,s)+r}. \tag{2.4}$$

Taking the small noise limit, the solution reduces to a Gaussian measure with mean $\mu^+(\cdot)$ and covariance function $\kappa^+(\cdot, \cdot)$ given by

$$\mu^+(\cdot) := \lim_{r\to 0} \mu(\cdot) = \frac{yk(s,\cdot)}{k(s,s)}, \qquad \kappa^+(\cdot,\cdot) := \lim_{r\to 0} \kappa(\cdot,\cdot) = k(\cdot,\cdot) - \frac{k(s,\cdot)k(\cdot,s)}{k(s,s)}. \tag{2.5}$$

Inserting $k(\tau, t) = \sigma \min\{\tau, t\}$, for any $t \in [0, T]$, it holds that

$$\mu^+(t) = y\frac{\min\{t,s\}}{s}, \qquad \kappa^+(t,t) = \sigma t - \sigma\frac{(\min\{t,s\})^2}{s}. \tag{2.6}$$

A particular case of $\mu^+(t)$ and $\kappa^+(t, t)$ is plotted in figure 2.1. The key property of this example is the continuum of hidden states not directly connected to the data. A posterior on a infinite dimensional space is being constrained by a finite dimensional observation, and the underlying problem is severely underdetermined. Draws from the posterior need to be constrained outside of the data space. This puts significant weight on the importance of the prior. The above example considered a linear observation model $h(x) = \theta x$, but many real-world observation models are nonlinear, e.g. the well known cubic sensor problem with $h(x) = \theta x + x^3$, (Steinberg et al., 1988). Such nonlinear observation models introduce ambiguity into the inference process. The considered example was also univariate, whereas many real world systems are highly multidimensional, with some geophysical applications having state-space dimensions in the millions (van Leeuwen, 2010). This is why the prior is so important in continuous-time systems and why it becomes more and more important as the frequency of observations decreases or the level of observation noise increases.

2.3 Stochastic processes

The chosen prior model in this thesis is a diffusion process. Subsumed by the class of Markov processes, diffusion processes have additional desirable qualities. This section outlines the basics of the theory of diffusion processes described in terms of Itô stochastic differential equations. It is important to fully characterise the properties of the diffusion process to prepare them for use in variational calculus. Some of the standard results in the literature for Markov processes are reiterated here for reference in later sections.

2.3.1 Stochastic differential equations

Given a deterministic initial condition $\mathbf{x}_0$ and a vector-valued function $\mathbf{f}(\mathbf{x}, t)$, the differential equation $\frac{d\mathbf{x}}{dt}(t) = \mathbf{f}(\mathbf{x}, t)$ can be considered short hand notation for the integral equation $\mathbf{x}(t) = \mathbf{x}_0 + \int_0^t \mathbf{f}(\mathbf{x}, t)$. This integral equation describes the time evolution of a deterministic system with state-vector $\mathbf{x}(t)$. To inject stochasticity into the system, the Wiener process $\mathbf{w}_t$ taking values in $\mathbb{R}^{d_\mathbf{x}}$ is used, and how, where, and when this stochasticity is injected is controlled by a matrix-valued function $\mathbf{Q}(\mathbf{x}, t)$. For any random vector initial condition $\boldsymbol{\xi}$ with $\left\langle ||\boldsymbol{\xi}||_2 \right\rangle < \infty$, such a system is described by a stochastic integral equation

$$\mathbf{x}_t = \boldsymbol{\xi} + \int_0^t \mathbf{f}(\mathbf{x}_s, s) ds + \int_0^t \mathbf{Q}(\mathbf{x}_s, s) d\mathbf{w}_s, \quad t \in [0, T], \tag{2.7}$$

which will often be written in short hand notation as a stochastic differential equation

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)dt + \mathbf{Q}(\mathbf{x}_t, t)d\mathbf{w}_t, \quad \mathbf{x}_0 = \boldsymbol{\xi}. \tag{2.8}$$

The following theorem characterises some simple constraints that can be placed on $\mathbf{f}(\mathbf{x}_t, t)$ and $\mathbf{Q}(\mathbf{x}_t, t)$ so that the resulting process $\mathbf{x}_t$ from equation (2.7) is continuous and unique. These are desirable qualities to avoid explosive behaviour and ambiguity that can hinder inference.

Theorem 2.1. *Let $\mathbf{f}(\mathbf{x}, t)$ and $\mathbf{Q}(\mathbf{x}, t)$ denote vector-valued and matrix-valued functions, respectively, measurable on $\mathbb{R}^{d_\mathbf{x}} \times [0, T]$. Assume there exists a constant $K > 0$ such that, for all $t \in [0, T]$ and $\mathbf{x}, \mathbf{z} \in \mathbb{R}^{d_\mathbf{x}}$:*

1. *(Lipschitz condition)*

$$||\mathbf{Q}(\mathbf{x},t)-\mathbf{Q}(\mathbf{z},t)||_F \leq K||\mathbf{x}-\mathbf{z}||_2, \qquad ||\mathbf{f}(\mathbf{x},t)-\mathbf{f}(\mathbf{z},t)||_2 \leq K||\mathbf{x}-\mathbf{z}||_2; \quad (2.9)$$

2. *(Linear growth condition)*

$$||\mathbf{Q}(\mathbf{x},t)||_F^2 \leq K(1+||\mathbf{x}||^2), \qquad ||\mathbf{f}(\mathbf{x},t)||_2 \leq K(1+||\mathbf{x}||^2). \quad (2.10)$$

Then for any measurable random vector $\boldsymbol{\xi}$ *with* $\langle||\boldsymbol{\xi}||_2\rangle < \infty$*, the stochastic integral equation*

$$\mathbf{x}_t = \boldsymbol{\xi} + \int_0^t \mathbf{f}(\mathbf{x}_s, s)ds + \int_0^t \mathbf{Q}(\mathbf{x}_s, s)d\mathbf{w}_s, \quad t \in [0,T], \quad (2.11)$$

has a unique continuous solution.

Most SDE models used in applications in general will meet the requirements in theorem (2.1). Though it is simple to construct examples that do not by defining functions $\mathbf{f}(\mathbf{x},t)$ such that the deterministic equation $\mathbf{x}(t) = \mathbf{x}_0 + \int_0^t \mathbf{f}(\mathbf{x},t)$ is not unique or well defined at some $t \in [0,T]$.

2.3.2 Markov processes

Definition 2.2. *Any* $\mathbb{R}^{d_\mathbf{x}}$*-valued stochastic process* $\{\mathbf{x}_t\}_{t\in[0,T]}$*, is said to satisfy the* Markov property *if for any* $0 \leq t_0 < t_1 < \cdots t_k < t \leq T$*, the equality*

$$P(\mathbf{x}_t \in B|\mathbf{x}_{t_0} \ldots \mathbf{x}_{t_k}) = P(\mathbf{x}_t \in B|\mathbf{x}_{t_k}) \quad (2.12)$$

holds for all Borel sets $B \subseteq \mathbb{R}^{d_\mathbf{x}}$*. Any stochastic process* $\{\mathbf{x}_t\}_{t\in[0,T]}$*, is said to be a* Markov process *if it satisfies the Markov property.*

For a realisation $\mathbf{x}'$ of $\mathbf{x}_s$, $s < t$, $P(\mathbf{x}_t \in B|\mathbf{x}_s = \mathbf{x}')$ defines a probability measure on the σ-algebra $\mathcal{B}$ of Borel subsets of $\mathbb{R}^{d_\mathbf{x}}$ such that

$$P(\mathbf{x}_t \in B|\mathbf{x}_s) \int_B p(\mathbf{x}, t; \mathbf{x}', s)d\mathbf{x} \quad (2.13)$$

for all Borel sets $B \in \mathcal{B}$. The quantity $p(\mathbf{x}, t; \mathbf{x}', s)$ is referred to as the *transition density*. Although, in general, insufficient for the specification of the probability law of the process, the forms $p(\mathbf{x}_t)$, $p(\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k})$, and $p(\mathbf{x}_t|\mathbf{x}_s)$ will be used to denote the marginal, joint marginal, and transition densities, respectively, as is common in the filtering literature (Jazwinski, 1970, Maybeck, 1979). The marginal $p(\mathbf{x}_t) = (\mathbf{x}, t)$ is considered a function of time t and a realisation $\mathbf{x}$ of $\mathbf{x}_t$, and similarly for the joint density $p(\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k})$. Importantly, from the definition of a

Markov process $\{\mathbf{x}_t\}_{t\in[0,T]}$, for any $0 < t_1 < \cdots t_k \leq T$, it holds that

$$p(\mathbf{x}_0, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_k}) = p(\mathbf{x}_0) \prod_{i=1}^{k} p(\mathbf{x}_{t_i}|\mathbf{x}_{t_{i-1}}). \tag{2.14}$$

For any $0 \leq t_1 < t_2 < t_3 \leq T$, the transition densities $p(\mathbf{x}_t|\mathbf{x}_s)$ satisfy the *Chapman-Kolmogorov* equation

$$p(\mathbf{x}_{t_3}|\mathbf{x}_{t_1}) = \int p(\mathbf{x}_{t_3}|\mathbf{x}_{t_2}) p(\mathbf{x}_{t_2}|\mathbf{x}_{t_1}) d\mathbf{x}_{t_2}. \tag{2.15}$$

In summary, the key property of a Markov process is the independence of its future from its past given the present. This type of conditional structure is natural for the temporal models considered here. While higher order models with dependencies that stretch beyond the present are discussed at certain points, they are not in the scope of the core content of the thesis. The decomposition of the joint distribution in equation (2.14), and the propagation step in equation (2.15), form two of the key components of a recursive Bayesian approach to inference in Markov processes. importantly, the unique continuous solution to the stochastic integral equation in theorem 2.1 is a Markov process (Kuo, 2006, Theorem 10.6.2). Therefore the above Markov methods provide the required tools for performing inference and other probabilistic operations in models defined using stochastic differential equations that meet the requirements of the theorem.

2.3.3 Diffusion processes

Diffusion process models are used universally to describe many continuous-time systems encountered in computational neuroscience (Feng, 2004), financial mathematics (Kabanov et al., 2006), physics and chemistry (van Kampen, 2007), biology (Wilkinson, 2006), image processing and computer vision (Weickert, 2001). For any $\mathbb{R}^{d_\mathbf{x}}$-valued vector $\mathbf{x}$, let $x^{(i)}$ denote its i^{th} component, and let $||\mathbf{x}||_2$ denote the squared norm of $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$. The notation $\epsilon \downarrow 0$ implies ϵ tends to zero from above, and is used when it is important that ϵ remains positive.

Definition 2.3. *Any $\mathbb{R}^{d_\mathbf{x}}$-valued Markov process $\{\mathbf{x}_t\}_{t\in[0,T]}$, is said to be a* diffusion process *if its transition density $p(\mathbf{z}, t; \mathbf{x}, s)$ satisfies the following conditions for any $t \in [0, T]$, $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$:*

$$\begin{aligned} 0 &= \lim_{\epsilon\downarrow 0} \frac{1}{\epsilon} \int_{||\mathbf{x}-\mathbf{z}||_2 \geq c} p(\mathbf{z}, t+\epsilon; \mathbf{x}, s) d\mathbf{z}, \quad \forall c > 0 & (2.16)\\ g^{(i)}(\mathbf{x}, t) &= \lim_{\epsilon\downarrow 0} \frac{1}{\epsilon} \int_{\mathbb{R}^{d_\mathbf{x}}} (z^{(i)} - x^{(i)}) p(\mathbf{z}, t+\epsilon; \mathbf{x}, s) d\mathbf{z} & (2.17)\\ D^{(ij)}(\mathbf{x}, t) &= \lim_{\epsilon\downarrow 0} \frac{1}{\epsilon} \int_{\mathbb{R}^{d_\mathbf{x}}} (z^{(i)} - x^{(i)})(z^{(j)} - x^{(j)}) p(\mathbf{z}, t+\epsilon; \mathbf{x}, s) d\mathbf{z}. & (2.18) \end{aligned}$$

The limits $g^{(i)}(\mathbf{x}, t)$ and $D^{(ij)}(\mathbf{x}, t)$ must exist for the definition to hold. Condition (2.16) ensures that the stochastic process $\mathbf{x}_t$ does not have instantaneous jumps. It implies the transition density reduces to the Dirac function $\delta(\mathbf{z} - \mathbf{x})$ as $\epsilon \downarrow 0$. The integrals in conditions (2.17) and

(2.18) are in fact defined for all Borel set in $\mathbb{R}^{d_{\mathbf{x}}}$. It is easy to show the limits $g^{(i)}(\mathbf{x},t)$ and $D^{(ij)}(\mathbf{x},t)$ are independent of the size of each Borel set, and therefore defined as above.

Definition 2.4. *The vector* $\mathbf{g}(\mathbf{x},t) = \big(g^{(i)}(\mathbf{x},t)\big)_{i\in\mathbb{N}_{d_{\mathbf{x}}}}$ *and the matrix* $\mathbf{D}(\mathbf{x},t) = \big(D^{(ij)}(\mathbf{x},t)\big)_{i,j\in\mathbb{N}_{d_{\mathbf{x}}}}$ *define the* drift *and the* diffusion matrix*, respectively, of the diffusion process* $\mathbf{x}_t$*.*

An alternative formulation for $\mathbf{g}(\mathbf{x},t)$ and $\mathbf{D}(\mathbf{x},t)$, which will be utilised later, is given by

$$\mathbf{g}(\mathbf{x},t) = \lim_{\epsilon\downarrow 0}\frac{1}{\epsilon}\big\langle \mathbf{x}_{t+\epsilon}-\mathbf{x}\big\rangle \tag{2.19}$$

$$\mathbf{D}(\mathbf{x},t) = \lim_{\epsilon\downarrow 0}\frac{1}{\epsilon}\big\langle (\mathbf{x}_{t+\epsilon}-\mathbf{x})(\mathbf{x}_{t+\epsilon}-\mathbf{x})^{\mathrm{T}}\big\rangle. \tag{2.20}$$

Thus, we can loosely consider $\mathbf{g}(\mathbf{x},t)$ and $\mathbf{D}(\mathbf{x},t)$, respectively, as the rate of change of the mean and the covariance of $\mathbf{x}_t$ at position $\mathbf{x}$ and time t. It follows directly from equation (2.20) that $\mathbf{D}(\mathbf{x},t)$ is a positive definite matrix for all $\mathbf{x}$ and t. Using definition 2.3 the following theorem holds.

Theorem 2.5. *Let* $\mathbf{f}(\mathbf{x},t)$ *and* $\mathbf{Q}(\mathbf{x},t)$ *be functions specified in theorem 2.1. Assume that* $\mathbf{f}(\mathbf{x},t)$ *and* $\mathbf{Q}(\mathbf{x},t)$ *are continuous on* $\mathbb{R}^{d_{\mathbf{x}}}\times[0,T]$*. Then the solution* $\mathbf{x}_t$ *given by equation* (2.11) *is a diffusion process with drift* $\mathbf{g}(\mathbf{x},t)=\mathbf{f}(\mathbf{x},t)$ *and diffusion matrix* $\mathbf{D}(\mathbf{x},t)=\mathbf{Q}(\mathbf{x},t)\mathbf{Q}^{\mathrm{T}}(\mathbf{x},t)$*.*

This theorem provides a sound method for constructing diffusion process priors in Itô SDE form, by simply ensuring that $\mathbf{g}(\mathbf{x},t)$ and $\sqrt{\mathbf{D}}(\mathbf{x},t)$ meet the conditions of theorem 2.1 and are continuous on $\mathbb{R}^{d_{\mathbf{x}}}\times[0,T]$, where $\sqrt{\mathbf{A}}$ denotes the square root of positive definite matrix $\mathbf{A}$. While various applications will not meet this criterion, all prior processes in the thesis from this point on are assumed to meet these continuity conditions as well as the Lipschitz and linear growth conditions in theorem 2.1. The same assumption cannot be made for the drift of the posterior. For low values of observation noise, the posterior drift will have to make abrupt changes to represent observations that are close in time but spatial distant. It is therefore important to fully characterise the properties of a vector valued function $\mathbf{g}(\mathbf{x},t)$ such that it defines the drift of a diffusion process $\mathbf{x}_t$.

2.3.4 Fokker-Planck equation

The Fokker-Planck equation provides us with a means for obtaining marginal and transition densities $p(\mathbf{x}_t)$ and $p(\mathbf{x}_t|\mathbf{x}_s), t > s$, respectively, from a diffusion process in Itô SDE form. Many stochastic models in science are described through the use of Itô SDEs. The Fokker-Planck or Kolmogorov forward equation is the necessary tool for converting these models into forms that are compatible with probabilistic inference. While this cannot always be done in closed form, the Fokker-Planck equation generates a vector field over the space of probabilities that can direct us towards solutions and, most importantly, can be integrated into a variational formulation of inference. Consider the following Itô SDE

$$d\mathbf{x}_t = \mathbf{g}(\mathbf{x}_t, t)dt + \sqrt{\mathbf{D}}(\mathbf{x}_t, t)d\mathbf{w}_t, \quad t \in [0, T], \tag{2.21}$$

where $\mathbf{g}(\mathbf{x}, t)$ and $\mathbf{D}(\mathbf{x}, t)$ are continuous functions of $\mathbf{x}$ and t, satisfy the Lipschitz and growth conditions in theorem 2.1, and $\mathbf{D}(\mathbf{x}, t)$ is positive definite. It follows that $\sqrt{\mathbf{D}}(\mathbf{x}_t, t)$ also satisfies the conditions in theorem 2.1 and equation (2.21) has a unique continuous solution. It is a Markov process, and therefore characterised by the density function $p(\mathbf{x}_t) = p(\mathbf{x}, t)$ for all $t \in [0, T]$ and the transition probability density function $p(\mathbf{x}_t|\mathbf{x}_s) = p(\mathbf{x}, t; \mathbf{x}', s)$ for all $t > s \in [0, T]$. The assumption of continuity makes equation (2.21) a diffusion process and its transition densities $p(\mathbf{x}_t|\mathbf{x}_s), t > s$ can be characterised in the following way. Assuming $p(\mathbf{x}, t)$ is differentiable in t, let ∂_t denote the time derivative operator

$$\partial_t[p(\mathbf{x}, t)] = \frac{\partial p(\mathbf{x}, t)}{\partial t}. \tag{2.22}$$

Assuming $\mathbf{g}(\mathbf{x}, t)$, $p(\mathbf{x}, t)$, and $\mathbf{D}(\mathbf{x}, t)$ are once differentiable in $\mathbf{x}$ and $p(\mathbf{x}, t)$ and $\mathbf{D}(\mathbf{x}, t)$ are twice differentiable in $\mathbf{x}$, let $\vec{\mathcal{K}}_{\mathbf{g}}$ denote the partial-differential Fokker-Planck operator

$$\vec{\mathcal{K}}_{\mathbf{g}}[p(\mathbf{x}, t)] = -\nabla^{\mathrm{T}}\left[p(\mathbf{x}, t)\mathbf{g}(\mathbf{x}, t)\right] + \frac{1}{2}\mathrm{tr}\left\{\left(\nabla\nabla^{\mathrm{T}}\right)\left[\mathbf{D}(\mathbf{x}, t)p(\mathbf{x}, t)\right]\right\} \tag{2.23}$$

where ∇ is the vector differential operator. The two operators defined in (2.22) and (2.23) can be coupled together to describe the interplay between temporal and spatial variations of the densities $p(\mathbf{x}, t)$ and $p(\mathbf{x}, t; \mathbf{x}', s)$. Indeed, under the above assumptions, $p(\mathbf{x}, t)$ and $p(\mathbf{x}, t; \mathbf{x}', s)$ satisfy the Fokker-Planck equations (Jazwinski, 1970)

$$\begin{aligned} \left(\partial_t - \vec{\mathcal{K}}_{\mathbf{g}}\right)[p(\mathbf{x}, t)] &= 0 & (2.24) \\ \left(\partial_t - \vec{\mathcal{K}}_{\mathbf{g}}\right)[p(\mathbf{x}, t; \mathbf{x}', s)] &= 0. & (2.25) \end{aligned}$$

Equations (2.24) and (2.25) describe how $p(\mathbf{x}, t)$ and $p(\mathbf{x}, t; \mathbf{x}', s)$ evolve over time given initial conditions $p(\mathbf{x}_0)$ and $\delta(\mathbf{x} - \mathbf{x}')$ respectively. The first is deterministic. The second is stochastic due to its dependency on the value $\mathbf{x}'$ taken by $\mathbf{x}_s$. Computing $p(\mathbf{x}_t|\mathbf{x}_0)$ using equation (2.25) and propagating $p(\mathbf{x}_0)$ through $p(\mathbf{x}_t|\mathbf{x}_0)$ using marginalisation, also produces $p(\mathbf{x}_t)$. It is simple to convert a continuous-time model to a discrete-time one using equation (2.25), allowing for the use of recursive Bayesian methods to perform inference and assimilate data. Unfortunately, Kolmogorov's forward equation is solvable in only a few simple cases. Thus, even before the assimilation of data has been considered, there are significant computational issues in capturing the probabilistic quantities of a continuous-time model. Despite not having a closed form, it is important to characterise when the Fokker-Planck equation has a well defined solution.

Theorem 2.6. *Let* $\mathbf{g}(\mathbf{x},t)$ *and* $\mathbf{D}(\mathbf{x},t)$ *be continuous in* t *and* $\mathbf{x}$*, and let* $\mathbf{g}(\mathbf{x},t)$*,* $\mathbf{D}(\mathbf{x},t)$*,* $\mathbf{g}_{\mathbf{x}}(\mathbf{x},t)$*,* $\mathbf{D}_{\mathbf{x}}(\mathbf{x},t)$*, and* $\mathbf{D}_{\mathbf{xx}}(\mathbf{x},t)$ *satisfy the Lipschitz and linear growth conditions in theorem 2.1. Then equation* (2.25) *has a unique solution satisfying the initial condition* $\delta(\mathbf{x}-\mathbf{x}')$.

2.3.4.1 Kolmogorov backward equation

Another important operator is the adjoint of the Fokker-Planck, or Kolmogorov backward, operator $\overleftarrow{\mathcal{K}}_{\mathbf{g}}$. For any twice differentiable function $\phi(\mathbf{x})$,

$$\overleftarrow{\mathcal{K}}_{\mathbf{g}}[\phi(\mathbf{x})] := \mathbf{g}^{\mathrm{T}}(\mathbf{x},t)\nabla[\phi(\mathbf{x})] + \frac{1}{2}\mathrm{tr}\Big\{\mathbf{D}(\mathbf{x},t)\big(\nabla\nabla^{\mathrm{T}}\big)[\phi(\mathbf{x})]\Big\}. \tag{2.26}$$

Note that equation (2.23) can also be defined for functions that are simply functions of $\mathbf{x}$ and not dependent on t. Importantly, the transition density $p(\mathbf{x}', s; \mathbf{x}, t)$, for $s > t$, satisfies the Kolmogorov backward equation

$$\Big(\partial_t + \overleftarrow{\mathcal{K}}_{\mathbf{g}}\Big)[p(\mathbf{x}', s; \mathbf{x}, t)] = 0. \tag{2.27}$$

Equation (2.25) describes the evolution of the transition density with respect to the future variables with the past variables fixed. Equation (2.27) describes the evolution of the transition density with respect to the past variables with the future variables fixed.

2.4 Partially observed diffusion processes

2.4.1 Latent Markov process

Consider a prior Markov process $\{\mathbf{x}_t\}_{t\in[0,T]}$, a finite set of observation times $\mathcal{T} \subset [0,T]$, and an observation model $\{p(\mathbf{y}|\mathbf{x},t_i)\}_{t_i\in\mathcal{T}}$. When $\{\mathbf{x}_t\}_{t\in[0,T]}$ is a diffusion process this model describes a partially observed diffusion process. While the observations can be continuous in time, only discrete time observations are considered here. The optimal smoothing problem equates to estimating the conditional probabilities $\{p(\mathbf{x}_t|\mathbf{Y})\}_{t\in[0,T]}$. This can be done by decomposing $p(\mathbf{x}_t|\mathbf{Y})$ into the product of two independent filters. First let us define the two following sets,

$$\mathbf{Y}_{\le t} \;=\; \{\mathbf{y}_i | t_i \le t\} \tag{2.28}$$

$$\mathbf{Y}_{> t} \;=\; \{\mathbf{y}_i | t_i > t\}. \tag{2.29}$$

The joint density $p(\mathbf{x}_t, \mathbf{Y})$ can be decomposed in two different ways, given by

$$p(\mathbf{x}_t, \mathbf{Y}) \;=\; p(\mathbf{x}_t|\mathbf{Y})p(\mathbf{Y}_{>t}|\mathbf{Y}_{\le t})p(\mathbf{Y}_{\le t}) \tag{2.30}$$

$$p(\mathbf{x}_t, \mathbf{Y}) \;=\; p(\mathbf{Y}_{>t}|\mathbf{x}_t, \mathbf{Y}_{\le t})p(\mathbf{x}_t|\mathbf{Y}_{\le t})p(\mathbf{Y}_{\le t}). \tag{2.31}$$

Combining equations (2.30) and (2.31), and using the fact that $\mathbf{Y}_{>t}$ is independent of $\mathbf{Y}_{\leq t}$ given $\mathbf{x}_t$, yields

$$p(\mathbf{x}_t|\mathbf{Y}) = \frac{p(\mathbf{Y}_{>t}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{Y}_{\leq t})}{p(\mathbf{Y}_{>t}|\mathbf{Y}_{\leq t})}. \tag{2.32}$$

This is the symmetric form of the smoothing density $p(\mathbf{x}_t|\mathbf{Y})$. The quantity $p(\mathbf{x}_t|\mathbf{Y}_{\leq t})$ is referred to as the optimal filter; the probability of $\mathbf{x}_t$ given the past data $\mathbf{Y}_{\leq t}$ up to time t, and the quantity $p(\mathbf{Y}_{>t}|\mathbf{x}_t)$ is referred to as the likelihood; the probability of the future data $\mathbf{Y}_{>t}$ after time t given $\mathbf{x}_t$. If $p(\mathbf{x}_t|\mathbf{Y}_{\leq t})$ and $p(\mathbf{Y}_{>t}|\mathbf{x}_t)$ are known, then $p(\mathbf{Y}_{>t}|\mathbf{Y}_{\leq t})$ can be computed simply by marginalisation,

$$p(\mathbf{Y}_{>t}|\mathbf{Y}_{\leq t}) = \int p(\mathbf{Y}_{>t}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{Y}_{\leq t})d\mathbf{x}_t, \tag{2.33}$$

and the optimal smoothing problem is complete, with $p(\mathbf{x}_t|\mathbf{Y})$ given by equation (2.32). The independence of $p(\mathbf{x}_t|\mathbf{Y}_{\leq t})$ and $p(\mathbf{Y}_{>t}|\mathbf{x}_t)$ from future and past observations, respectively, allows them to be computed in a simple sequential fashion. When $\{\mathbf{x}_t\}_{t\in[0,T]}$ is a diffusion process, this is done through the use of the Kolmogorov forward and backward equations.

2.4.2 Forward filter

Let $\{\mathbf{x}_t\}_{t\in[0,T]}$ denote a partially observed diffusion process with drift $\mathbf{f}(\mathbf{x},t)$ and diffusion matrix $\mathbf{D}(\mathbf{x},t)$ satisfying the conditions in theorem 2.6. Let $p_F(\mathbf{x},t) \equiv p(\mathbf{x}_t|\mathbf{Y}_{\leq t})$ denote the optimal filter density, generated by conditioning $\mathbf{x}_t$ on $\mathbf{Y}_{\leq t}$, with initial condition $p_F(\mathbf{x},0) = p(\mathbf{x}_0)$. Let t_i^- denote a time infinitesimally close to t_i and assume $|\mathcal{T}| = m$. Finding $p_F(\mathbf{x},t)$ is referred to as the optimal filtering problem. The optimal filtering problem is solved as follows:

1. For $i = 1, \dots, m$, repeat:

   (a) *Prediction:* In the interval $[t_{i-1}, t_i^-)$, solve the initial value problem

   $$\frac{\partial q(\mathbf{x},t)}{\partial t} = \vec{\mathcal{K}}_{\mathbf{f}}[q(\mathbf{x},t)], \quad q(\mathbf{x},t_{i-1}) = p_F(\mathbf{x},t_{i-1}). \tag{2.34}$$

   Define $p_F(\mathbf{x},t) = q(\mathbf{x},t)$ for all $t \in [t_{i-1}, t_i^-)$.

   (b) *Update:* At the observation time t_i, use Bayes' rule to compute

   $$p_F(\mathbf{x},t_i) = \frac{p(\mathbf{y}_i|\mathbf{x},t_i)p_F(\mathbf{x},t_i^-)}{p(\mathbf{y}_i|\mathbf{Y}_{<t_i})}. \tag{2.35}$$

Step (b) is often referred to as the forward jump condition. In theory, the above algorithm generates an optimal filter $p_F(\mathbf{x},t)$ for all times $t \in [0,T]$.

2.4.3 Information filter

Let $\{\mathbf{x}_t\}_{t\in[0,T]}$ denote a partially observed diffusion process with drift $\mathbf{f}(\mathbf{x},t)$ and diffusion matrix $\mathbf{D}(\mathbf{x},t)$ satisfying the conditions in theorem 2.6. Let us define $\psi(\mathbf{x},t)$ such that

$$\psi(\mathbf{x},t) := \frac{p(\mathbf{x}_t|\mathbf{Y})}{p_F(\mathbf{x},t)} = \frac{p(\mathbf{Y}_{>t}|\mathbf{x}_t)}{p(\mathbf{Y}_{>t}|\mathbf{Y}_{\leq t})}. \tag{2.36}$$

We see that $\psi(\mathbf{x},t)$ is composed of the likelihood $p(\mathbf{Y}_{>t}|\mathbf{x}_t)$ and the normalisation constant $p(\mathbf{Y}_{>t}|\mathbf{Y}_{\leq t})$. One of the benefits of this definition is the resulting end condition $\psi(\mathbf{x},T)=1$, which is independent of $p_F(\mathbf{x},T)$. Using equations (2.36) and (2.35), and the fact that $p(\mathbf{x}_t|\mathbf{Y})$ is differentiable in t, yields

$$\psi(\mathbf{x},t_i^-) \overset{(2.36)}{=} \frac{p(\mathbf{x}_{t_i}|\mathbf{Y})}{p_F(\mathbf{x},t_i^-)} \overset{(2.35)}{=} \frac{p(\mathbf{y}_i|\mathbf{x},t_i)p(\mathbf{x}_{t_i}|\mathbf{Y})}{p(\mathbf{y}_i|\mathbf{Y}_{<t_i})p_F(\mathbf{x},t_i)} \overset{(2.36)}{=} \frac{p(\mathbf{y}_i|\mathbf{x},t_i)\psi(\mathbf{x},t_i)}{p(\mathbf{y}_i|\mathbf{Y}_{<t_i})}. \tag{2.37}$$

Thus $\psi(\mathbf{x},t)$ satisfies a backward jump condition in analogy to equation (2.35). We also have the following proposition:

Proposition 2.7. *Between observations times, $\psi(\mathbf{x},t)$ satisfies the backward Kolmogorov equation*

$$\left(\partial_t + \overleftarrow{\mathcal{K}}_{\mathbf{g}}\right)[\psi(\mathbf{x},t)] = 0. \tag{2.38}$$

Proof. Given the discrete time nature of the observations, for any time t such that $t_{i-1} < t < t_i$, and for any time step $\delta t > 0$ such that $t+\delta t < t_i$, it holds that

$$\begin{aligned}\psi(\mathbf{x},t) &= \frac{p(\mathbf{Y}_{>t}|\mathbf{x},t)}{p(\mathbf{Y}_{>t}|\mathbf{Y}_{\leq t})} = \frac{p(\mathbf{Y}_{>t+\delta t}|\mathbf{x},t)}{p(\mathbf{Y}_{>t+\delta t}|\mathbf{Y}_{\leq t+\delta t})} &(2.39)\\ &= \int d\mathbf{x}' \frac{p(\mathbf{Y}_{>t+\delta t}|\mathbf{x}',t+\delta t)}{p(\mathbf{Y}_{>t+\delta t}|\mathbf{Y}_{\leq t+\delta t})} p(\mathbf{x}',t+\delta t;\mathbf{x},t) &(2.40)\\ &= \int d\mathbf{x}'\psi(\mathbf{x}',t+\delta t)p(\mathbf{x}',t+\delta t;\mathbf{x},t), &(2.41)\end{aligned}$$

where $p(\mathbf{x}',t+\delta t;\mathbf{x},t) = p(\mathbf{x}'_{t+\delta t}|\mathbf{x}_t)$ denotes the transition density. Applying the operator $(\partial_t + \overleftarrow{\mathcal{K}}_{\mathbf{g}})$ to the $(\mathbf{x},t)$ variables on both sides of equation (2.41), moving the differentials inside the integral, and noting that $p(\mathbf{x}',t+\delta t;\mathbf{x},t)$ satisfies the backward Kolmogorov equation (2.27) with respect to $(\mathbf{x},t)$, yields

$$\left(\partial_t + \overleftarrow{\mathcal{K}}_{\mathbf{g}}\right)[\psi(\mathbf{x},t)] = \int d\mathbf{x}'\psi(\mathbf{x}',t+\delta t)\left(\partial_t + \overleftarrow{\mathcal{K}}_{\mathbf{g}}\right)[p(\mathbf{x}',t+\delta t;\mathbf{x},t)] = 0. \tag{2.42}$$

Therefore $\psi(\mathbf{x},t)$ also satisfies the backward Kolmogorov equation. □

Finding $\psi(\mathbf{x},t)$ will be referred to as the information filter problem, in analogy to the discrete time problem (Briers et al., 2010). Recall the end condition $\psi(\mathbf{x},T)=1$. The information filter is computed as follows:

1. For $i = m, \dots, 1$, repeat:

   (a) *Retrodiction:* In the interval $[t_i, t_{i+1}^-)$, solve the end value problem

$$\frac{\partial q(\mathbf{x},t)}{\partial t} = -\overleftarrow{\mathcal{K}}_{\mathbf{f}}[q(\mathbf{x},t)], \quad q(\mathbf{x}, t_{i+1}^-) = \psi(\mathbf{x}, t_{i+1}^-). \tag{2.43}$$

   Define $\psi(\mathbf{x},t) = q(\mathbf{x},t)$ for all $t \in [t_{i-1}, t_i^-)$.

   (b) *Downdate:* At the observation time t_i, compute

$$\psi(\mathbf{x}, t_i^-) = \frac{p(\mathbf{y}_i|\mathbf{x}, t_i)\psi(\mathbf{x}, t_i)}{p(\mathbf{y}_i|\mathbf{Y}_{<t_i})}. \tag{2.44}$$

Equation (2.43) is integrated backwards in time from t_{i+1}^- to t_i. In theory, the above algorithm generates an information filter $\psi(\mathbf{x},t)$ for all times $t \in [0,T]$.

2.4.4 Posterior drift filter

After computing the optimal filter $p_F(\mathbf{x},t)$ in section 2.4.2 and the information filter $\psi(\mathbf{x},t)$ in section 2.4.3, the smoothing density $p_s(\mathbf{x},t) = p(\mathbf{x}_t|\mathbf{Y})$ is given simply by $p_s(\mathbf{x},t) = p_F(\mathbf{x},t)\psi(\mathbf{x},t)$. We have the following important result.

Theorem 2.8. *The smoothing density $p_s(\mathbf{x},t) = p(\mathbf{x}_t|\mathbf{Y})$ satisfies the Fokker-Planck equation*

$$\frac{\partial p_s}{\partial t} = \overrightarrow{\mathcal{K}}_{\mathbf{g}}[p_s], \quad \mathbf{g}(\mathbf{x},t) = \mathbf{f}(\mathbf{x},t) + \mathbf{D}(\mathbf{x},t)\nabla \ln \psi(\mathbf{x},t). \tag{2.45}$$

Thus, exact inference boils down to computing computing the information filter $\psi(\mathbf{x},t)$ backwards in time using the steps in section 2.4.3 and computing the smoothing density $p_s(\mathbf{x},t)$ forward in time by solving the Kolmogorov forward equation in (2.45) with drift $\mathbf{g}(\mathbf{x},t)$. This backward-forward algorithm is an alternative to the symmetric approach of the previous section.

Proof. It holds that

$$\begin{align}
\frac{\partial p_s}{\partial t} &= \frac{\partial p_F}{\partial t}\psi - p_F\frac{\partial \psi}{\partial dt} \tag{2.46}\\
&= \overrightarrow{\mathcal{K}}_{\mathbf{f}}[p_F]\psi - p_F\overleftarrow{\mathcal{K}}_{\mathbf{f}}[\psi] \tag{2.47}\\
&= \Big[\tfrac{1}{2}\mathrm{tr}\big((\nabla\nabla^{\mathrm{T}})(\mathbf{D}p_F)\big) - \nabla^{\mathrm{T}}(p_F\mathbf{f})\Big]\psi - p_F\Big[\tfrac{1}{2}\mathrm{tr}\big(\mathbf{D}(\nabla\nabla^{\mathrm{T}})\psi\big) + \mathbf{f}^{\mathrm{T}}\nabla\psi\Big] \tag{2.48}\\
&= -\nabla^{\mathrm{T}}(p_s\mathbf{f}) + \tfrac{1}{2}\Big[\mathrm{tr}\big(\psi(\nabla\nabla^{\mathrm{T}})(p_F\mathbf{D})\big) - \mathrm{tr}\big(p_F\mathbf{D}(\nabla\nabla^{\mathrm{T}})\psi\big)\Big] \tag{2.49}\\
&= -\nabla^{\mathrm{T}}(p_s\mathbf{f}) + \tfrac{1}{2}\Big[\mathrm{tr}\big(\psi(\nabla\nabla^{\mathrm{T}})(p_F\mathbf{D})\big) + 2\nabla^{\mathrm{T}}\mathbf{D}p_F\nabla\psi \nonumber\\
&\quad +\mathrm{tr}\big(p_F\mathbf{D}(\nabla\nabla^{\mathrm{T}})\psi\big)\Big] - \Big[\nabla^{\mathrm{T}}\mathbf{D}p_F\nabla\psi + \mathrm{tr}\big(p_F\mathbf{D}(\nabla\nabla^{\mathrm{T}})\psi\big)\Big] \tag{2.50}\\
&= -\nabla^{\mathrm{T}}(p_s\mathbf{f}) + \tfrac{1}{2}\mathrm{tr}\big((\nabla\nabla^{\mathrm{T}})(\mathbf{D}p_s)\big) - \nabla^{\mathrm{T}}\mathbf{D}p_F\nabla\psi \tag{2.51}\\
&= -\nabla^{\mathrm{T}}(p_s(\mathbf{f} + \mathbf{D}\nabla\ln\psi)) + \tfrac{1}{2}\mathrm{tr}\big((\nabla\nabla^{\mathrm{T}})(\mathbf{D}p_s)\big) \tag{2.52}\\
&= \overrightarrow{\mathcal{K}}_{\mathbf{g}}[p_s] \tag{2.53}
\end{align}$$

where

$$\mathbf{g}(\mathbf{x},t) = \mathbf{f}(\mathbf{x},t) + \mathbf{D}(\mathbf{x},t)\nabla \ln \psi(\mathbf{x},t). \tag{2.54}$$

□

The posterior drift $\mathbf{g}(\mathbf{x},t)$ in equation (2.45) is a very important quantity. It highlights how exact and approximate inference can be phrased as learning a posterior drift, as opposed to learning the explicit measure. It has very close connections to control theory, and the idea of posterior drift permeates all throughout the thesis. The approach to exact inference described in theorem (2.8) makes perfect sense in the control setting; the posterior inherits the diffusive nature of the prior and the likelihood backwards pass collects all the information required to guide diffusions towards the data. It would be very important to characterise when equation (2.45) has well defined unique solutions. It is likely that nonlinear observation operators will introduce ambiguity and generate solutions that are not unique.

2.5 Approximate inference

While it is possible, in theory, to perform the backward-forward or two-filter approaches to exact inference and obtain a continuous time solution, this is generally not viable in practice. It is possible to solve the filtering and smoothing problems discussed in section 2.4 up to a finite set of times, by discretising time and converting the continuous time problem to a discrete time one. Indeed, let $0 = t_0 < \cdots < t_N = T$ and $\delta_i = t_{i+1} - t_i$ denote a discretisation of time, and define $\mathbf{x}_i = \mathbf{x}(t_i)$, $\mathbf{X} = (\mathbf{x}_0, \ldots, \mathbf{x}_N)$ and $\mathbf{Y} = \{\mathbf{y}_i, i \in \mathcal{I} \subseteq [0:N]\}$. Consider the latent Markov model with joint distribution

$$p(\mathbf{X},\mathbf{Y}) = p(\mathbf{x}_0)\prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{x}_{i-1}) \prod_{j\in\mathcal{I}} p(\mathbf{y}_j|\mathbf{x}_j). \tag{2.55}$$

A visual representation of the probabilistic structure of this model is shown in figure 2.2. To convert the continuous-time problem into a discrete-time problem, only the transition densities $p(\mathbf{x}_i|\mathbf{x}_{i-1})$ need to be found. While this can be done exactly using equation (2.25), this equation is solvable in only a small number of cases. For this reason, the standard approach in many inference schemes is to apply a first-order Euler-Maruyama discretisation (Kloeden & Platen, 1992) to (2.21),

$$\begin{aligned} t_{i+1} &= t_i + \delta_i & (2.56)\\ \mathbf{x}_{i+1} &= \mathbf{x}_i + \mathbf{g}(\mathbf{x}_i, i)\delta_i + \sqrt{\delta_i \mathbf{D}}(\mathbf{x}_i, i)\boldsymbol{\xi}_i & (2.57) \end{aligned}$$

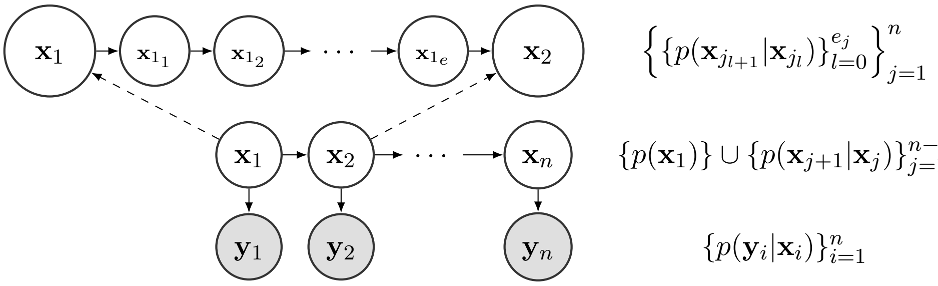

$\mathbf{x}_1$
$\mathbf{x}_{1_1}$
$\mathbf{x}_{1_2}$
$\cdots$
$\mathbf{x}_{1_e}$
$\mathbf{x}_2$
$\left\{\{p(\mathbf{x}_{j_{l+1}}|\mathbf{x}_{j_l})\}_{l=0}^{e_j}\right\}_{j=1}^{n}$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\cdots$
$\mathbf{x}_n$
$\{p(\mathbf{x}_1)\} \cup \{p(\mathbf{x}_{j+1}|\mathbf{x}_j)\}_{j=1}^{n-1}$
$\mathbf{y}_1$
$\mathbf{y}_2$
$\mathbf{y}_n$
$\{p(\mathbf{y}_i|\mathbf{x}_i)\}_{i=1}^{n}$

FIGURE 2.2: Hidden Markov model with intermediary hidden states. The figure consists of (from left to right): the corresponding Bayesian network; the probabilities needed to complete the model quantitatively for each corresponding (horizontal) level of hierarchy. The dashed arrows corresponds to a "zoom" into the upper network, showing the intermediary hidden-states lying between $\mathbf{x}_1$ and $\mathbf{x}_2$.

where $\boldsymbol{\xi}_i$ is a Gaussian distributed random variable with zero-mean and variance δ_i for $i = 0, \ldots, N-1$. This approximation leads to a transition density of the form

$$p(\mathbf{x}_{i+1}|\mathbf{x}_i) \propto \frac{\varphi\Big(\big(\delta_i \mathbf{D}(\mathbf{x}_i, i)\big)^{-\frac{1}{2}}\big(\mathbf{x}_{i+1} - \mathbf{x}_i - \delta_i \mathbf{g}(\mathbf{x}_i, i)\big)\Big)}{|\delta_i \mathbf{D}(\mathbf{x}_i, i)|^{\frac{1}{2}}} \tag{2.58}$$

where $\varphi(\mathbf{x}) = (2\pi)^{-d_{\mathbf{x}}/2} \exp(-\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}}{2})$. The posterior $p(\mathbf{X}|\mathbf{Y})$ can be decomposed (Briers et al., 2004) into

$$p(\mathbf{X}|\mathbf{Y}) = p(\mathbf{x}_0|\mathbf{Y}) \prod_{j=1}^{N} p(\mathbf{x}_{j+1}|\mathbf{x}_j, \mathbf{Y}) \tag{2.59}$$

where $\{p(\mathbf{x}_{j+1}|\mathbf{x}_j, \mathbf{Y})\}_{i=1}^{N}$ and $p(\mathbf{x}_0|\mathbf{Y})$ can be considered a set of posterior transition densities and a posterior initial condition, respectively. In analogy to the continuous-time setting, algorithms that attempt to learn the set of conditional densities $\{p(\mathbf{x}_i|\mathbf{Y})\}_{i=1}^{n}$ are referred to as smoothing algorithms and algorithms that attempt to learn the set of conditional densities $\{p(\mathbf{x}_i|\mathbf{Y}_{\leq i})\}_{i=1}^{n}$, are referred to as filtering algorithms, where $\mathbf{Y}_{\leq i} := (\mathbf{y}_1, \ldots, \mathbf{y}_i)$. With $p(\mathbf{x}_0)$ defining an initial density, the problem of discrete-time filtering reduces to the simple forward algorithm 1. Once the filtering densities have been obtained, letting $p(\mathbf{x}_N|\mathbf{Y})$ define the end condition, the discrete-time smoothing densities can be built iteratively using the backward algorithm 2. There is no conditioning on data points because the empirical information has already been incorporated into beliefs in the forward sweep. The backward sweep simply passes the

```
Algorithm: forward algorithm
for $i = 1, \ldots, N$ do
    Marginalisation step; $p(\mathbf{x}_i|\mathbf{Y}_{\leq i}) = \int p(\mathbf{x}_i|\mathbf{x}_{i-1}) p(\mathbf{x}_{i-1}|\mathbf{Y}_{\leq i-1}) d\mathbf{x}_{i-1}$.
    if $i \in \mathcal{I}$ then do
        Conditioning step; $p(\mathbf{x}_i|\mathbf{Y}_{\leq i}) = \frac{1}{p(\mathbf{Y}_i|\mathbf{Y}_{\leq i-1})} p(\mathbf{Y}_i|\mathbf{x}_i) p(\mathbf{x}_i|\mathbf{Y}_{\leq i-1})$.
```

Algorithm 1: Forward algorithm for discrete-time state-space models or first-order hidden Markov models.

```
Algorithm: backward algorithm
for $i = N-1, \ldots, 0$ do
    Marginalisation step; $p(\mathbf{x}_i|\mathbf{Y}) = \int p(\mathbf{x}_i|\mathbf{x}_{i+1}, \mathbf{Y}) p(\mathbf{x}_{i+1}|\mathbf{Y}) d\mathbf{x}_{i-1}$.
```

Algorithm 2: Backward algorithm for discrete-time state-space models. Requires the forward algorithm to compute the backward transition $p(\mathbf{x}_i|\mathbf{x}_{i+1}, \mathbf{Y})$.

information in the data points back to all the hidden states occurring before the time of each observation. This is done in the backward algorithm through the use of the transition density $p(\mathbf{x}_i|\mathbf{x}_{i+1}, \mathbf{Y})$ given by

$$p(\mathbf{x}_i|\mathbf{x}_{i+1}, \mathbf{Y}) = \frac{p(\mathbf{x}_i|\mathbf{Y}_{\leq i}) p(\mathbf{x}_{i+1}|\mathbf{x}_i)}{p(\mathbf{x}_{i+1}|\mathbf{Y}_{\leq i})}. \tag{2.60}$$

In analogy to the continuous-time setting, exact filtering requires one forward sweep of the data. Exact smoothing algorithms on the other hand require one forward and one backward sweep. For the general case, the marginalisation and conditioning steps in both algorithms require approximation. This leads to a multitude of recursive approximate inference schemes that deal with these sequential updates in a local fashion.

2.5.1 Recursive Monte Carlo

General particle-filters are recursive Monte Carlo methods that transform the recursive estimation problem into a stochastic system of interacting particles. Assuming $p(\mathbf{x}_i|\mathbf{Y}_{\leq i})$ is approximated by a set of particles $\{\mathbf{x}_i^{(j)}\}_{j=1}^n$, the bootstrap filter (Gordon et al., 1993) propagates the current particle set forward in time to construct a new set $\{\mathbf{x}_{i+1}^{(j)}\}_{j=1}^n$ according to

$$\mathbf{x}_{i+1}^{(j)} = \mathbf{x}_i^{(j)} + \mathbf{g}(\mathbf{x}_i^{(j)}, i)\delta_i + \mathbf{D}(\mathbf{x}_i^{(j)}, i)\boldsymbol{\xi}_i^{(j)} \tag{2.61}$$

for $j = 1, \ldots, n$. The new set $\{\mathbf{x}_{i+1}^{(j)}\}_{j=1}^n$ then constitutes an empirical approximation of the propagated density $p(\mathbf{x}_{i+1}|\mathbf{Y}_{\leq i})$. The same method is used in the regularised particle filter (Doucet et al., 2001) and the auxiliary particle filter (Pitt & Shephard, 1999), but with additional resampling and lookahead components, respectively. Accompanying the particles is a set of weights $\{w_i^j\}_{j=1}^n$ which allow the importance of each particle, under the data, to be controlled using Bayes rule. On the backward pass, these weights ignore the data points and are updated using the backward transition function $p(\mathbf{x}_j|\mathbf{x}_{j+1}, \mathbf{Y})$ in (2.60), which requires a closed form for the forward transition density $p(\mathbf{x}_{i+1}|\mathbf{x}_i)$. For the discretisation scheme in (2.57) a conditionally Gaussian closed form for $p(\mathbf{x}_{i+1}|\mathbf{x}_i)$ is given in (2.58), but this forces the smoothing method to use a first-order linear approximation. A more general approximation scheme is given by

$$t_{i+1} = t_i + \delta_i \tag{2.62}$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \mathrm{RK}(\mathbf{g}(\mathbf{x}_i, i), \mathbf{D}(\mathbf{x}_i, i), \boldsymbol{\xi}_i, \delta_i) \tag{2.63}$$

where $\mathrm{RK}(\cdot,\cdot,\cdot,\cdot)$ can be one of many stochastic Runge-Kutta step-functions (Murray & Storkey, 2011). These general schemes involve nonlinear maps of the noise variable $\boldsymbol{\xi}_i$, and therefore in most cases do not admit a closed form for $p(\mathbf{x}_{i+1}|\mathbf{x}_i)$. By mapping realisations of $\mathbf{x}_i^{(j)}$ and $\boldsymbol{\xi}_i$ through the Runge-Kutta step-function it is still possible to propagate samples $\mathbf{x}_i^{(j)} \mapsto \mathbf{x}_{i+1}^{(j)}$ forward in time under this scheme, so the above filtering methods work in this setting, but the reweighting procedure in the backward pass for smoothing is not possible without a closed form for $p(\mathbf{x}_{i+1}|\mathbf{x}_i)$. Therefore most smoothing methods cannot be used in conjunction with higher order numerical integration schemes. This is remedied in the work of Murray & Storkey (2011), by using of a proposal that includes the transition density and cancels it out from the corresponding update equations. There is a distinction between discrete-time and continuous-time methods in the context of particle filtering and smoothing, where discrete-time methods are forced to use linear numerical integration schemes such as (2.57) and continuous-time methods have the luxury of higher-order nonlinear schemes. The label "continuous-time method" is due to the fact that the algorithmic solutions (i.e. the paths of the particles) can be considered continuous-time processes, to the extent that they can be realised by any desired stochastic numerical-integration scheme. The benefit of this is that higher order schemes facilitate the use of larger and adaptive time-step-sizes while maintaining accuracy, leading to improved computational efficiency.

2.5.2 Assumed density smoothing

When the drift $\mathbf{f}(\mathbf{x}, t)$ and observation operator $\mathbf{h}(\mathbf{x}, t)$ are linear, all densities involved in the filtering and smoothing problems are Gaussian. The forward algorithm then equates to the classical discrete-time Kalman filter of Kalman (1960), and the backward algorithm equates the Rauch-Tung-Striebel smoother of Rauch et al. (1965). These follow from the fundamental rules for marginalising and conditioning Gaussian distributions. See also Kschischang et al. (2001) for a more general message passing characterisation. When $\mathbf{f}(\mathbf{x}, t)$ and $\mathbf{h}(\mathbf{x}, t)$ are nonlinear, the required marginal beliefs have either no closed-form, or the descriptions of the beliefs grow so quickly over time as to become computationally intractable. An idea, originating in Kushner (1967), is to project the marginal beliefs back onto a tractable family after each stage of inference. For the forward algorithm, the resulting algorithm is referred to as the assumed density filter (ADF), or the Gaussian filter for the particular case of the Gaussian family (Maybeck, 1979). These methods attempt to retain as much of the structure of the Kalman filtering and smoothing algorithms as possible, often to a fault. The most recent formulations focus on the assumed-Gaussian setting (Ito, 2000, Särkkä & Hartikainen, 2010a) and examine efficient methods for approximating the intractable integrals involved in the projection step. Once the projection step is dealt with, the rules for marginalisation and conditioning Gaussians are all applicable. While this simple formulation makes ADF and assumed density smoothing (ADS) appealing, they are inherently restricted to one forward and one backward pass of the data. This is adequate when inference is exact, but when inference is approximate, multiple forward and backward iterations are required (Heskes & Zoeter, 2002). Recent work from Särkkä (2007, 2010), Särkkä & Sarmavouri (2011) has shown how continuous-time limits can be taken of ADF and ADS solutions to the discretised problem, with approximate transition probability (2.58). This has lead to a highly stable and fast assumed Gaussian smoothing algorithm and, in a way, counteracts the approximation error from using a first-order linear discretisation. The method leads to a set of backward differential equations (DE) for the mean and covariance of the Gaussian approximation, which can then be solved using any applicable DE solver. While experiments show that ADS is not good at representing nonlinearities, such as transitions between stable equilibrium points, it shows good results in approximately linear regions. In many systems, trajectories commonly inhabit such regions.

2.5.3 Expectation propagation

Expectation propagation (Minka, 2001) is a popular technique in approximate inference and easily adapted to the discrete-time smoothing problem. This is the setting considered in Heskes & Zoeter (2002), Yu et al. (2007, 2006), Zoeter & Heskes (2005). Let $\tau_i(\mathbf{x}_i, \mathbf{x}_{i-1})$ denote the site function

$$\tau_i(\mathbf{x}_i, \mathbf{x}_{i-1}) := \begin{cases} p(\mathbf{x}_i|\mathbf{x}_{i-1})p(\mathbf{y}_i|\mathbf{x}_i) & i \in \mathcal{I} \\ p(\mathbf{x}_i|\mathbf{x}_{i-1}) & i \notin \mathcal{I} \\ p(\mathbf{x}_0) & i = 0. \end{cases} \tag{2.64}$$

Then the joint distribution in equation (2.55) can be written

$$p(\mathbf{X}, \mathbf{Y}) = \prod_{i=0}^{N} \tau_i(\mathbf{x}_i, \mathbf{x}_{i-1}). \tag{2.65}$$

The primary aim of EP is to approximate the marginals of the smoothing density $p(\mathbf{X}|\mathbf{Y})$ and its normalisation constant $Z := \int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})d\mathbf{X}$. It provides a very general smoothing algorithm that isn't constrained to one forward and backward pass. Under any message passing scheme (Heskes & Zoeter, 2002), a fully factorised EP approximation $q(\mathbf{X})$ will equate to the product of $N+1$ forward and backward messages $\hat{\alpha}_i$ and $\hat{\beta}_i$, such that

$$q(\mathbf{X}) = \prod_{i=0}^{N} q_i(\mathbf{x}_i) = \prod_{i=0}^{N} \frac{1}{Z_i}\hat{\alpha}_i(\mathbf{x}_i)\hat{\beta}_i(\mathbf{x}_i). \tag{2.66}$$

Iterating through the i's in any order desired, the EP updates are given in algorithm 3. The scheduling method generally used for smoothing consists of a forward pass - updating the $\hat{\alpha}_i$'s while keeping the $\hat{\beta}_i$'s fixed, and a backward pass - updating the $\hat{\beta}_i$'s while keeping the $\hat{\alpha}_i$'s fixed. The $\sim$ symbol indicates the updates may or may not occur. The projection step consists of minimising a distance measure $\mathcal{D}(p, q)$, usually the relative entropy between p and q, over a tractable class $\mathcal{Q}$. It simplifies to a moment matching condition $q_i(\mathbf{x}_i, \mathbf{x}_{i-1}) \overset{\text{MM}}{\leftrightarrow} \tilde{p}_i(\mathbf{x}_i, \mathbf{x}_{i-1})$ when $\mathcal{Q}$ is an exponential family. The main problem in EP-based smoothing is dealing with the

```
Algorithm: Chain-EP algorithm
repeat
    Choose; $i \in \mathbb{N}_N$
    Build; $\tilde{p}_i(\mathbf{x}_i, \mathbf{x}_{i-1}) = \frac{1}{\tilde{Z}_i} \hat{\alpha}_{i-1}(\mathbf{x}_{i-1}) t_i(\mathbf{x}_i, \mathbf{x}_{i-1}) \hat{\beta}_i(\mathbf{x}_i)$
    Project; $q_i(\mathbf{x}_i, \mathbf{x}_{i-1}) = \arg\min_{q \in \mathcal{Q}} \mathcal{D}(\tilde{p}_i, q)$
    $\sim$ Update; $\hat{\alpha}_i^{\text{new}}(\mathbf{x}_i) = Z_i \frac{\int q_i(\mathbf{x}_i, \mathbf{x}_{i-1}) d\mathbf{x}_{i-1}}{\hat{\beta}_i(\mathbf{x}_i)}$
    $\sim$ Update; $\hat{\beta}_{i-1}^{\text{new}}(\mathbf{x}_{i-1}) = Z_{i-1} \frac{\int q_i(\mathbf{x}_i, \mathbf{x}_{i-1}) d\mathbf{x}_i}{\hat{\alpha}_{i-1}(\mathbf{x}_{i-1})}$
until covergence
```

Algorithm 3: General EP algorithm for linear Markov chain.

projection step, or moment matching for exponential families, in analogous fashion to the main problem in assumed density smoothing. This is remedied through various numerical approximations in the references given at the start of the section. EP-smoothing and ADF are in fact very closely related, intersecting if only one pass of EP is considered. But while ADF and ADS are restricted to one forward and one backward pass, EP allows for multiple passes which do not, necessarily, have to respect the temporal ordering of the data. This is due to EP's symmetrical representation of the smoothing density as the product of forward and backward messages.

2.5.4 Hybrid Monte-Carlo

Using the first-order Euler-Maruyama scheme in (2.57), the joint distribution in equation (2.55) can be written

$$p(\mathbf{X}, \mathbf{Y}) \propto \exp\Big(- \mathcal{H}(\mathbf{X}, \mathbf{Y}) \Big), \tag{2.67}$$

where $\mathcal{H} = \mathcal{U}_0 + \mathcal{H}_{dyn} + \mathcal{H}_{obs}$ is a discretised Hamiltonian such that

$$\begin{aligned}
\mathcal{U}_0 &= -\log p(\mathbf{x}_0) && (2.68)\\
\mathcal{H}_{dyn} &= \sum_{i=0}^{N-1} \frac{\delta t}{2} \left[\frac{\mathbf{x}_{i+1} - \mathbf{x}_i}{\delta t} - \mathbf{f}(\mathbf{x}_i, i) \right]^{\mathrm{T}} \mathbf{D}^{-1}(\mathbf{x}_i, i) \left[\frac{\mathbf{x}_{i+1} - \mathbf{x}_i}{\delta t} - \mathbf{f}(\mathbf{x}_i, i) \right] && (2.69)\\
\mathcal{H}_{obs} &= \sum_{i \in \mathcal{I}} \frac{1}{2} \big[\mathbf{h}(\mathbf{x}_i, i) - \mathbf{y}_i \big]^{\mathrm{T}} \mathbf{R}^{-1}(\mathbf{x}_i, i) \big[\mathbf{h}(\mathbf{x}_i, i) - \mathbf{y}_i \big]. && (2.70)
\end{aligned}$$

While the Hamiltonian is in fact just the regularised log-likelihood of the data under the Euler-Maruyama approximation, the Hamiltonian approach recasts the joint distribution as a statistical mechanical system (Alexander et al., 2005). The dynamic component $\mathcal{H}_{obs}$ is related to the (discretised) path-integral formulation of the marginal densities given in (Risken, 1996, Section 4.4.2). This global formulation, as opposed to the local formulation of recursive estimation, leads to two main approaches for approximate inference, both rooted in statistical mechanics. The observations in (2.67) are fixed (realised), therefore it can be considered a density on the hidden states $\mathbf{X} = (\mathbf{x}_0, \dots, \mathbf{x}_N)$, and the $\mathbf{Y}$ notation can be dropped. When normalised with respect to $\mathbf{X}$, equation (2.67) equates to the posterior $p(\mathbf{X}|\mathbf{Y})$. In Hybrid Monte-Carlo (HMC), the Hamiltonian of a canonical distribution is used to generate samples, which are then either accepted or rejected in a Metropolis-Hastings subloop. The algorithm generates a Markov chain $\mathbf{X}^{(k)} \rightarrow \mathbf{X}^{(k+1)}$ using a fictitious deterministic system

$$\frac{d\mathbf{X}}{d\tau} = \mathbf{P} \tag{2.71}$$

$$\frac{d\mathbf{P}}{d\tau} = -\nabla_{\mathbf{X}}\hat{\mathcal{H}}(\mathbf{X},\mathbf{P}) \tag{2.72}$$

where $\mathbf{P} = (\mathbf{p}_0, \ldots, \mathbf{p}_N)$ represent auxiliary momentum variables and $\hat{\mathcal{H}} = \mathcal{H}_{pot} + \mathcal{H}_{kin}$, such that

$$\mathcal{H}_{pot} = \mathcal{U}_0 + \mathcal{H}_{dyn} + \mathcal{H}_{obs} \tag{2.73}$$

$$\mathcal{H}_{kin} = \frac{1}{2}\sum_{i=0}^{N} \mathbf{p}_i^{\mathrm{T}} \mathbf{p}_i. \tag{2.74}$$

The potential $\mathcal{H}_{pot}$ represents the Hamiltonian of any probabilistic model, but has been specialised here to the discretised state-space model in equation (2.67). The system at each step of the Markov chain is initialised by setting $\mathbf{X}(\tau = 0) = \mathbf{X}^{(k)}$ and sampling $\mathbf{P}(\tau = 0)$ from a Gaussian distribution. The system is then integrated forward in time using one of a number of numerical methods tailored for Hamiltonian type equations in (2.71) and (2.72) (see Neal (2010) for more details). After a predefined amount of time $J\delta\tau$, the state $\mathbf{X}(\tau = J\delta\tau)$ is proposed as $\mathbf{X}^{(k+1)}$, and accepted with probability

$$\min\left\{1, \exp\left(\hat{\mathcal{H}}^{(k)} - \hat{\mathcal{H}}^{(k+1)}\right)\right\} \tag{2.75}$$

where $\hat{\mathcal{H}}^{(k)} = \hat{\mathcal{H}}(\mathbf{X}^{(k)}, \mathbf{P}^{(k)})$ and $\hat{\mathcal{H}}^{(k+1)} = \hat{\mathcal{H}}(\mathbf{X}^{(k+1)}, \mathbf{P}^{(k+1)})$. After sufficient burning and subsampling, the samples $\{\mathbf{X}^{(k)}\}_{k=1}^{M}$ represent draws from the posterior $p(\mathbf{X}|\mathbf{Y})$, and provide important properties such canonical invariance and ergodicity (Neal, 2010). The application of HMC to state-space models has been performed in Alexander et al. (2005) and Shen et al. (2010), though only to very simple models. The main problem with applying HMC is finding good initial conditions for the Markov chain, and dealing with highly nonlinear Hamiltonian functions. The formulation can be extended to higher order discretisation schemes than the Euler-Maruyama method considered here (Restrepo, 2008). The HMC algorithm can then, in a way, be considered a continuous-time algorithm, though a reduction in step size or the introduction of higher order schemes can affect the computationally complexity considerably. Special considerations need to be made when applying the HMC algorithm to a continuous-time model. The arbitrariness of the size of the discretisation step δt can result in a state vector $\mathbf{X}$ of arbitrarily large dimension. Therefore it helps to consider the algorithm in a infinite dimensional setting (Stuart, 2010). This ensures, in the limit, the algorithm behaves in a reasonable way. The use of a preconditioning matrix (Generalised Hybrid Monte Carlo) can reduce the correlation between the state vectors at different times, helping the HMc algorithm to be more efficient in that state-space model setting (Alexander et al., 2005).

2.5.5 Variational methods

Variational methods in machine learning are based upon minimising the Kullback-Leibler divergence or relative entropy between the approximation and the true posterior. In finite dimensions, for an approximation $q(\mathbf{X})$, the divergence between $q(\mathbf{X})$ and the posterior $p(\mathbf{X}|\mathbf{Y})$ in equation (2.1) is given by

$$\mathrm{KL}[q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})] = \int q(\mathbf{X}) \log \left(\frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \right). \tag{2.76}$$

Importantly, the Kullback-Leibler divergence is positive and zero if and only if $q(\mathbf{X}) = p(\mathbf{X}|\mathbf{Y})$ almost everywhere. Parametric variational methods equip $q(\mathbf{X};\boldsymbol{\theta})$ with a vector of parameters and optimise over the set of parameterised densities. Due to the presence of p, the divergence $\mathrm{KL}[q||p]$ can not be minimised directly, but using simple algebraic manipulations yields

$$-\log p(\mathbf{Y}) + \mathrm{KL}[q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})] \leq \mathcal{F}(q(\mathbf{X};\boldsymbol{\theta})) \tag{2.77}$$

where $\mathcal{F}(q(\mathbf{X};\boldsymbol{\theta}))$ denotes the free-energy

$$\mathcal{F}(q(\mathbf{X};\boldsymbol{\theta})) = \mathrm{KL}[q(\mathbf{X})||p(\mathbf{X})] - \big\langle \log p(\mathbf{Y}|\mathbf{X}) \big\rangle_{q(X)}. \tag{2.78}$$

Here $\langle \mathbf{x} \rangle$ denotes the expectation of a random variable $\mathbf{x}$ and $\langle \phi(\mathbf{x}) \rangle_{q(\mathbf{x})}$ denotes the expectation of $\phi(\mathbf{x})$ with respect to the density q. Given the KL divergence is positive and $p(\mathbf{Y})$ is constant with respect to q, minimising $\mathcal{F}(q(\mathbf{X};\boldsymbol{\theta}))$ over q or $\boldsymbol{\theta}$ is equivalent to minimising $\mathrm{KL}[q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})]$, and provides an upper bound on $-\log p(\mathbf{Y})$ which can be used for model selection.

The variational approach of Archambeau et al. (2007a), Archambeau & Opper (2011) assumes a prior drift $\mathbf{f}(\mathbf{x}, t)$ is provided to the SDE in equation (2.21). The scheme then attempts to learn an approximate drift $\mathbf{g}(\mathbf{x}, t)$ in the presence of data. These two drift functions generate SDEs that can be considered continuous-time limits of the Euler-Maruyama approximation in equation (2.57). For an arbitrary time-step δt, the discretisations $\mathbf{X} = (\mathbf{x}_0, \ldots, \mathbf{x}_N)$ fit naturally into the variational framework. The Euler-Maruyama scheme can be used to generate a discrete-time prior $p(\mathbf{X})$ using the drift $\mathbf{f}(\mathbf{x}, t)$ and a discrete-time approximation $q(\mathbf{X})$ using the drift $\mathbf{g}(\mathbf{x}, t)$. The Kullback-Leibler divergence between $p(\mathbf{X})$ and $q(\mathbf{X})$ is then given by

$$\mathrm{KL}[q(\mathbf{X})||p(\mathbf{X})] = \frac{\delta t}{2} \sum_{i=0}^{N-1} \left\langle ||\mathbf{g}(\mathbf{x}_i, i) - \mathbf{f}(\mathbf{x}, i)||^2_{\mathbf{D}(\mathbf{x}_i, i)} \right\rangle_{q_i} + \mathrm{KL}[q_0||p_0] \tag{2.79}$$

where $q_i(\mathbf{x}_i)$ and $p_i(\mathbf{x}_i)$ denote the marginals onto $\mathbf{x}_i$ of $q(\mathbf{X})$ and $p(\mathbf{X})$, respectively. Let P_{prior} and Q denote the limits of p and q as the time step tends to zero. Following the approach of

Archambeau et al. (2007a), taking the limit $\delta t \to 0$ in equation (2.79) yields

$$\mathrm{KL}[Q||P_{prior}] = \frac{1}{2}\int_0^T dt \Big\langle ||\mathbf{g}(\mathbf{x},t) - \mathbf{f}(\mathbf{x},t)||^2_{\mathbf{D}(\mathbf{x},t)} \Big\rangle_{q(\mathbf{x},t)} + \mathrm{KL}[q_0||p_0] \tag{2.80}$$

where $q(\mathbf{x},t)$ denotes the one-time marginal density of Q. Given the discrete-time nature of the observations in the state-space model, the log likelihood of the data is given by

$$\big\langle \log p(\mathbf{Y}|\mathbf{x}(\cdot)) \big\rangle_Q = \frac{1}{2}\int_0^T dt \sum_i \delta(t-t_i) \Big\langle \log p(\mathbf{y}_i|\mathbf{x}(t_i)) \Big\rangle_{q(\mathbf{x},t)}. \tag{2.81}$$

The variational smoothing algorithm then looks to minimise the *free-action*

$$\mathcal{F}(Q(\mathbf{x}(\cdot);\mathbf{g})) = \mathrm{KL}[Q||P_{prior}] - \big\langle \log p(\mathbf{Y}|\mathbf{x}(\cdot)) \big\rangle_Q \tag{2.82}$$

over g and Q. Importantly, this variational approach attempts to minimise the cumulated free-energy over time. The free-action $\mathcal{F}(Q(\mathbf{x}(\cdot);\mathbf{g}))$ bounds the accumulated log-evidence of the data, as opposed to bounding the log-evidence of the accumulated data as is done in many standard variational machine learning methods Jordan et al. (1999b).

2.6 Discussion

This chapter has introduced state-space models in the form of partially observed diffusion processes with discrete time observations, and the inference problem of assimilating data. Various approaches to learning the latent states of the system have been considered in a machine learning context. In most cases a Euler-Maruyama linear approximation is applied to the prior to deal with the intractable Kolmogorov equations. The Euler-Maruyama linear approximation has been shown to have undesirable qualities (Ozaki, 1993) and a good approximate inference scheme should not be restricted to using it. Monte Carlo algorithms remedy this problem through the use of higher order approximation schemes for the underlying SDE. This is done is a recursive setting in Murray & Storkey (2011) and in a path-integral setting in Restrepo (2008), though both methods are still in their infancy. With regards to deterministic methods, the assumed density (AD) smoothing algorithm of Särkkä & Sarmavouri (2011) is presented in a continuous-time form by taking the continuous time limit of solutions of a discrete-time AD algorithm applied to the Euler-Maruyama approximation of the prior. This allows for higher-order numerical integration methods to be applied to the resulting ordinary differential equations (ODEs). This algorithm is also in its infancy and it is not known if it can be extended beyond the Gaussian setting. Also in the deterministic setting, current EP-smoothing methods are only applicable in discrete-time, require a closed form for the transition density, and are also most commonly used in Gaussian form. Archambeau et al. (2007a), Archambeau & Opper (2011) use the variational approach of section (2.5.5) to generate a solution set of continuous-time ODEs. While the algorithm is formulated only for Gaussian approximations, it appears to be the strongest contender out of all the deterministic methods for approximate inference for smoothing in state-space models. This is due to various reasons:

- It has a natural formulation in continuous-time.
- It has a well defined objective function that allows for direct application of gradient methods and a well-defined notion of convergence.
- It fits seamlessly into a model selection and system identification framework.

As with all methods, the variational framework of Archambeau & Opper (2011) does not scale well with the size of the state-dimension, though it appears faster than MCMC methods in most situations (Shen et al., 2010). Its main weaknesses when compared to Monte Carlo methods are as follows:

- It is currently restricted to Gaussian approximations and therefore does not have the descriptive power of MC methods.
- Its implementation requires a significant amount of preparation and auxiliary functions when compared to, essentially "black-box", Monte Carlo methods.

It is apparent that there is an underdeveloped gap in the literature for non-Gaussian deterministic approximate inference. Due to the strengths and flexibility of the variational framework described in section (2.5.5), it is natural to focus on extending this framework to handle non-Gaussian approximations, as opposed to focusing on the alternative EP-smoothing and assumed density smoothing frameworks whose weaknesses have been exposed.

Independently, both Gaussian mixture models and exponential families have universal approximation abilities to any continuous density function (Mclachlan & Peel, 2000, Wainwright & Jordan, 2008, respectively). The Gaussian density sits at the intersection of both these families; being a Gaussian mixture model with only one component and an exponential family density with only linear and quadratic sufficient statistics. Therefore any progress made with the trusty Gaussian is relevant to both the more general settings. Due to its infancy, there are still at lot of unanswered questions concerning the optimal Gaussian solution of Archambeau et al. (2007a), and strengthening its theory and implementation methods can only facilitate progress in the more general setting.

Chapter 3

General framework

Contribution

- This chapter extends the variational state-space-model learning method to the general setting, i.e. beyond the Gaussian case. The generalisation requires explicits form for the drift in a stochastic differentiable equation and the corresponding one-time marginal densities, which are generally not available. It is shown how this requirement can be weakened and the marginal densities of the process can be projected down onto a tractable approximating class. This enables the use of more flexible relations between the variational drift and the approximate marginals in the variational smoothing algorithm. The projection has a natural moment-matching interpretation and this is used to derive a projected version of the well known Fokker-Planck equation.

3.1 Introduction

The variational state-space model learning scheme derived in Archambeau et al. (2007a) - and featuring in Archambeau et al. (2007b), Archambeau & Opper (2011), Vrettas et al. (2010), and Shen et al. (2010) - relies upon the Gaussian assumption. For many nonlinear state-space models it is desirable for approximations to capture higher order moments than those of the Gaussian. This chapter discusses extending the variational approach of Archambeau et al. (2007a) to higher order distributions characterised by their moments. The first part of the chapter generalises the variational smoothing framework of Archambeau & Opper (2011). The chapter begins by considering the exact Bayesian solution as the optimal variational solution, and goes on to introduce a generalised variational smoothing algorithm with weakened constraints. Rather than re-derive the Gaussian case for the reader and then generalise, the section starts at the deep end; formulating the general case and using the Gaussian case as an example. The generalisation requires an explicit form for the drift of a diffusion process and the corresponding marginal densities, which are generally not available. While is relatively simple to define a tractable class of marginal densities or to fix a particular form of drift, finding a tractable pairing can be highly non-trivial, accept in the linear case. Therefore the explicit drift-marginal pairing requirement would make the generalised algorithm not easily usable in practice. For an arbitrary drift, marginal densities can quickly propagate into computationally intractable regions of the space of probability density functions. This propagation is dominated by the Fokker-Planck equation. Explicit drift-marginal pairings, are pairs of drifts and marginal densities that satisfy the Fokker-Planck equation. Luckily this requirement can be weakened and the marginal densities can be projected down onto a tractable approximating class. This allows for a mismatch between the variational drift and the approximate marginal densities in the variational smoothing algorithm, where the approximate marginal densities are projections of the marginal densities generated by the variational drift. This projection has a natural moment-matching interpretation and section 3.5.2 uses this to derive a projected version of the Fokker-Planck equation. While almost any drift can have its marginal densities projected onto a tractable class, it is recommended that the approximate pairings should be chosen to reduce the amount of information lost in the projection. In summary, the variational smoothing framework of Archambeau & Opper (2011) is generalised to general drift-marginal pairings. The strict requirement of an explicit drift-marginal pairing is weakened by allowing marginal densities that drift into intractable regions to be projected back onto a specified class. These projected marginal densities can be used as surrogates for computing marginal density expectations required in the variational smoothing algorithm by gradient based optimisation methods.

3.2 Optimal smoothing

The variational framework subsumes exact inference, and the choice of a Gaussian approximation - or any other suboptimal approximation - corresponds to replacing general constraints in the variational formulation with simplified moment constraints. The variational framework relies upon the free-energy formulation of learning outlined below.

3.2.1 Free-energy of a partially observed diffusion process

Let Q and P_{prior} denote probability measures on the space of paths generated by (2.21) with drifts $\mathbf{g}(\mathbf{x}, t)$ and $\mathbf{f}(\mathbf{x}, t)$ respectively. Let $q(\mathbf{x}, t)$ denote the one-time-slice marginal density of Q at time t. Define

$$E_{sde}(t) = \frac{1}{2}\Big\langle ||\mathbf{g}(\mathbf{x}, t) - \mathbf{f}(\mathbf{x}, t)||^2_{\mathbf{D}(\mathbf{x},t)} \Big\rangle_{q(\mathbf{x},t)}, \tag{3.1}$$

where $||\mathbf{x}||_{\mathbf{R}}$ denotes the Mahalanobis norm of $\mathbf{x}$ with respect to a positive-definite matrix $\mathbf{R}$, and $\langle \phi(\mathbf{x}) \rangle_{q(\mathbf{x})}$ denotes the expectation of $\phi(\mathbf{x})$ with respect to a density $q(\mathbf{x})$. From Archambeau et al. (2007a), the relative entropy $\mathrm{KL}[Q||P_{prior}]$ between Q and P_{prior}, is given by the integral

$$\mathrm{KL}[Q||P_{prior}] = \int_0^T E_{sde}(t)dt. \tag{3.2}$$

Let $\mathbf{Y} = (\mathbf{y}_{t_1}, \dots, \mathbf{y}_{t_m})$, $m \in \mathbb{N}$, denote a set of observations where each $\mathbf{y}_{t_i} \in \mathbb{R}^{d_\mathbf{y}}$ is observed at time $t_i \in [0, T]$. Define

$$E_{obs}(t) = \frac{1}{2} \sum_i \delta(t - t_i) \Big\langle ||\mathbf{y}_i - \mathbf{h}(\mathbf{x}, t_i)||^2_{\mathbf{R}(\mathbf{x}, t_i)} \Big\rangle_{q(\mathbf{x},t)} \tag{3.3}$$

where $\mathbf{h} : \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}_+ \to \mathbb{R}^{d_\mathbf{y}}$ is a nonlinear observation mapping and $\mathbf{R} : \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}_+ \to \mathbb{R}^{d_\mathbf{y} \times d_\mathbf{y}}$ is a positive-definite noise matrix. The expected negative-log-likelihood of the data, denoted $\mathcal{L}(Q)$, is given by the integral

$$\mathcal{L}(Q) = \int_0^T E_{obs}(t) dt. \tag{3.4}$$

Posterior approximations in this framework are characterised by their drift function $\mathbf{g}$ and represented by the corresponding measure Q. The free-energy $\mathcal{F}(Q)$ of an approximation Q, is defined as the sum of the relative entropy $\mathrm{KL}[Q||P_{prior}]$, where $\mathbf{f}$ denotes a prior-drift defined before the data is observed, and the expected negative-log-likelihood of the data $\mathcal{L}(Q)$, such that

$$\mathcal{F}(Q) = \mathcal{L}(Q) + \mathrm{KL}[Q||P_{prior}]. \tag{3.5}$$

Given that Q is characterised by its drift $\mathbf{g}(\mathbf{x}, t)$, it is reasonable to write the free-energy $\mathcal{F}(\mathbf{g})$ simply as a function of $\mathbf{g}(\mathbf{x}, t)$. Let P_{post} denote the posterior-measure obtained through exact Bayesian inference. As will be shown in the sequel, it holds that the global minimiser of the free-energy $\mathcal{F}(Q)$ is equivalent to the true posterior P_{post}. Furthermore, the one-time-slice marginal densities generated by the optimising-drift of $\mathcal{F}(\mathbf{g})$ are equivalent to the conditional (smoothing) densities $p(\mathbf{x}_t|\mathbf{Y})$ for all $t \in [0, T]$. Therefore the optimal Bayesian smoothing solution is characterised by a diffusion process with a posterior-drift. The posterior drift retains important properties of the prior $\mathbf{f}(\mathbf{x}, t)$, whilst guiding paths towards the data. The balance between prior and data depends strongly on the stochasticity of the system, characterised by $\mathbf{D}(\mathbf{x}, t)$ and $\mathbf{R}(\mathbf{x}, t)$.

3.2.2 Variational formulation of exact inference

Outlined in this section is the variational approach to exact inference, taken from Archambeau & Opper (2011) with a few minor adjustments. It makes explicit the incorporation of constraints relating the variational drift and to the corresponding variational marginal densities. It is these constraints that can be relaxed when constructing suboptimal variational approximations.

Let $\overrightarrow{\mathcal{K}}_\mathbf{g}[\cdot]$ and $\overleftarrow{\mathcal{K}}_\mathbf{g}[\cdot]$ denote, respectively, the Fokker-Planck forward operator and its adjoint

$$\begin{aligned} \overrightarrow{\mathcal{K}}_\mathbf{g}[p(\mathbf{x}, t)] &= -\nabla^{\mathrm{T}} [p(\mathbf{x}, t)\mathbf{g}(\mathbf{x}, t)] + \frac{1}{2}\mathrm{tr}\Big\{ (\nabla\nabla^{\mathrm{T}}) [\mathbf{D}(\mathbf{x}, t) p(\mathbf{x}, t)] \Big\} &(3.6)\\ \overleftarrow{\mathcal{K}}_\mathbf{g}[p(\mathbf{x}, t)] &= \mathbf{g}^{\mathrm{T}}(\mathbf{x}, t)\nabla[p(\mathbf{x}, t)] + \frac{1}{2}\mathrm{tr}\Big\{ \mathbf{D}(\mathbf{x}, t) (\nabla\nabla^{\mathrm{T}}) [p(\mathbf{x}, t)] \Big\}. &(3.7) \end{aligned}$$

Any drift function $\mathbf{g}(\mathbf{x}, t)$, arbitrary marginal density $q(\mathbf{x}, t)$, and dual variable $\rho(\mathbf{x}, t)$ satisfy the action functional

$$\int_0^T dt \Big\langle \big(\partial_t - \overrightarrow{\mathcal{K}}_{\mathbf{g}}\big)[q(\mathbf{x}, t)] \Big\rangle_{\rho(\mathbf{x},t)} = - \int_0^T dt \Big\langle \big(\partial_t + \overleftarrow{\mathcal{K}}_{\mathbf{g}}\big)[\rho(\mathbf{x}, t)] \Big\rangle_{q(\mathbf{x},t)} \tag{3.8}$$

where all expectations are over $\mathbf{x}$, and equality follows from $\overleftarrow{\mathcal{K}}_{\mathbf{g}}[\cdot]$ and $\overrightarrow{\mathcal{K}}_{\mathbf{g}}[\cdot]$ being dual-adjoint. Note that $\rho(\mathbf{x}, t)$ might not be a probability density and, in which case, expectations under $\rho(\mathbf{x}, t)$ are assumed to be finite integrals. Augmenting $\mathcal{F}(\mathbf{g})$ with (either of) (3.8), leads to the Lagrange functional

$$\mathcal{L}(\mathbf{g}, \rho) = \mathcal{F}(\mathbf{g}) - \int_0^T dt \Big\langle \big(\partial_t - \overrightarrow{\mathcal{K}}_{\mathbf{g}}\big)[q(\mathbf{x}, t)] \Big\rangle_{\rho(\mathbf{x},t)}. \tag{3.9}$$

Performing independent variations (see appendix A.2) of q and $\mathbf{g}$, and substituting $\rho(\mathbf{x}, t) = -\ln \psi(\mathbf{x}, t)$, leads to

$$\begin{aligned} U(\mathbf{x}, t)\psi(\mathbf{x}, t) &= \big(\partial_t + \overleftarrow{\mathcal{K}}_{\mathbf{g}}\big)[\psi(\mathbf{x}, t)] && (3.10) \\ \mathbf{g}(x, t) &= \mathbf{f}(x, t) + \mathbf{D}(\mathbf{x}, t)\nabla \ln \psi(\mathbf{x}, t) && (3.11) \end{aligned}$$

where

$$U(\mathbf{x}, t) = \frac{1}{2} \sum_i \delta(t - t_i) || \mathbf{y}_i - \mathbf{h}(\mathbf{x}, t_i) ||^2_{\mathbf{R}(\mathbf{x}, t_i)}. \tag{3.12}$$

Exact inference boils down to solving (3.10) backwards in time for $\psi(\mathbf{x}, t)$, with end condition $\psi(\mathbf{x}, T) = 1$. The Dirac functions in $u(\mathbf{x}, t)$ lead naturally to jumps for $\psi(\mathbf{x}, t)$ at the observation times t_i. Then the posterior drift given by $\mathbf{g}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}, t) + \mathbf{D}(\mathbf{x}, t)\nabla \ln \psi(\mathbf{x}, t)$ is used to solve (2.24) forwards in time to generate the smoothing densities $q(\mathbf{x}, t) = p(\mathbf{x}_t|\mathbf{Y})$ for all $t \in [0, T]$. This formulation matches the exact Bayesian solutions given in Eyink et al. (2004) and Archambeau & Opper (2011), which are derived through direct differentiation of the smoothing densities $p(\mathbf{x}_t|\mathbf{Y})$ (section 2.4.4).

The free-energy $\mathcal{F}(\mathbf{g})$ is solely a function of $\mathbf{g}(\mathbf{x}, t)$, but can only be written explicitly if $q(\mathbf{x}, t)$ is introduced to handle expectations. Ignoring for the moment that $q(\mathbf{x}, t)$ is the primary object of interest, the density $q(\mathbf{x}, t)$ can be considered an auxiliary variable, introduced to ease the optimisation problem of learning the optimal drift $\mathbf{g}(\mathbf{x}, t)$. To perform independent variations of q and $\mathbf{g}$, requires the incorporation of the constraint (3.8) to couple the variables together. It is this constraint that can be relaxed to allow for suboptimal variational approximations.

3.3 Generalised approximate inference

In this section a general formulation for variational inference is presented to deal with minimising the free energy integral in equation (3.5). The approach is novel by making no assumptions of linearity for the approximating drift and no assumption of Gaussianity for the approximating distribution. It deals with general marginal densities characterised by their vector of sufficient statistics and corresponding vector of moment parameters. This is done through the use of generalised moment equations that weaken the constraints used in the exact formulation of the previous section.

3.3.1 Optimal approximation

Let $\boldsymbol{\phi} : \mathbb{R}^{d_{\mathbf{x}}} \to \mathbb{R}^{d_{\gamma}}$ denote any twice-differentiable mapping $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_{d_\gamma}(\mathbf{x}))$. Starting from equation (2.24), using integration-by-parts it is simple to show that the drift function $\mathbf{g}(\mathbf{x}, t)$ and the marginal density $q(\mathbf{x}, t)$ satisfy the relation

$$\frac{\partial \langle \boldsymbol{\phi}(\mathbf{x}) \rangle_{q(\mathbf{x},t)}}{\partial t} - \left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}(\mathbf{x})] \right\rangle_{q(\mathbf{x},t)} = 0, \tag{3.13}$$

where the convention used for an operator $\mathcal{A}[\cdot]$ applied to a vector valued mapping $\boldsymbol{\phi}(\mathbf{x})$ is

$$\mathcal{A}[\boldsymbol{\phi}(\mathbf{x})] = \left(\mathcal{A}[\phi_1(\mathbf{x})], \ldots, \mathcal{A}[\phi_{d_\gamma}(\mathbf{x})]\right)^T. \tag{3.14}$$

It should be obvious that equation (3.13) is a weaker constraint than (2.24), inasmuch that the time and spatial variations of $q(\mathbf{x}, t)$ are coupled together only through projections of $q(\mathbf{x}, t)$ onto the component functions of $\boldsymbol{\phi}(\mathbf{x})$. When $\boldsymbol{\phi}(\mathbf{x})$ defines the vector of sufficient statistics of $q(\mathbf{x}, t)$ and $\boldsymbol{\gamma}(t) = \langle \boldsymbol{\phi}(\mathbf{x}, t) \rangle_{q(\mathbf{x},t)}$ defines its vector of (time-varying) moments, equation (3.13) leads naturally to a set of moment equations

$$\frac{\partial \boldsymbol{\gamma}(t)}{\partial t} = \left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}(\mathbf{x})] \right\rangle_{q(\mathbf{x},t)}. \tag{3.15}$$

Augmenting $\mathcal{F}(\mathbf{g})$ with the weaker constraint in (3.15) leads to a surrogate Lagrange functional

$$\hat{\mathcal{L}}(\mathbf{g}, \boldsymbol{\lambda}) = \mathcal{F}(\mathbf{g}) - \int_0^T dt \boldsymbol{\lambda}^{\mathrm{T}}(t) \left(\frac{\partial \boldsymbol{\gamma}(t)}{\partial t} - \left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}(\mathbf{x})] \right\rangle_{q(\mathbf{x},t)} \right) \tag{3.16}$$

where the dual variable $\boldsymbol{\lambda}(t)$ is a function of time and vector-valued to match the moment parameter $\boldsymbol{\gamma}(t)$. Assume that $\mathbf{g}(\cdot, t)$ is parameterised by a finite dimensional[1] vector $\boldsymbol{\pi}_t$. Taking

[1] This is a reasonable assumption, given that any viable approximate drift $\mathbf{g}(\mathbf{x}, t)$ will need to be representable using finite computational memory.

variations of (3.16) with respect to $\boldsymbol{\pi}_t$ and γ_t leads to a set of explicit gradients

$$\begin{aligned} \nabla_{\boldsymbol{\pi}_t}\hat{\mathcal{L}}(\mathbf{g},\boldsymbol{\lambda}) &= \nabla_{\boldsymbol{\pi}_t}E_{sde}(t) + \boldsymbol{\lambda}^{\mathrm{T}}(t)\nabla_{\boldsymbol{\pi}_t}\left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}(\mathbf{x})]\right\rangle_{q(\mathbf{x},t)} && (3.17)\\ \nabla_{\gamma_t}\hat{\mathcal{L}}(\mathbf{g},\boldsymbol{\lambda}) &= \nabla_{\gamma_t}E_{sde}(t) + \nabla_{\gamma_t}E_{obs}(t) + \frac{\partial\boldsymbol{\lambda}(t)}{\partial t} + \nabla_{\gamma_t}\left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}(\mathbf{x})]\right\rangle_{q(\mathbf{x},t)}^{\mathrm{T}}\boldsymbol{\lambda}(t). && (3.18)\end{aligned}$$

Setting (3.18) equal to zero and identifying the discrete time nature of $E_{obs}(t)$, leads to an adjoint equation

$$\frac{\partial\boldsymbol{\lambda}(t)}{\partial t} = -\nabla_{\gamma_t}E_{sde}(t) - \nabla_{\gamma_t}\left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}(\mathbf{x})]\right\rangle_{q(\mathbf{x},t)}^{\mathrm{T}}\boldsymbol{\lambda}(t) \tag{3.19}$$

with the jump conditions at the observations times $t_i \in \mathcal{T}$, given by

$$\boldsymbol{\lambda}(t_i) = \boldsymbol{\lambda}(t_i^+) - \nabla_{\gamma_t}E_{obs}(t_i), \tag{3.20}$$

where t^+ denotes a time infinitesimally close to (the right of) time t. Variational approximate-inference therefore equates to finding the global minimum of the surrogate Lagrangian $\hat{\mathcal{L}}(\mathbf{g},\boldsymbol{\lambda})$ with respect to $\boldsymbol{\pi}_t$, such that equations (3.15), (3.19), and (3.20) are satisfied. In algorithmic terms, a dynamic optimisation procedure is performed with gradient steps taken w.r.t. $\boldsymbol{\pi}_t$ and a sub-loop for solving equations (3.15), (3.19), and (3.20). The general variational smoothing algorithm can then be written:

[1] **Initialise:** $\boldsymbol{\pi}_t$ for all $t \in [0,T]$.

[2] **Solve:** *(forwards)* $t \in [0,T]$, $\gamma(0) = \gamma_0$,

$$\frac{\partial\gamma(t)}{\partial t} = \left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}(\mathbf{x})]\right\rangle_{q(\mathbf{x},t)}.$$

[3] **Solve:** *(backwards)* $t \in [0,T]$, $\boldsymbol{\lambda}(T) = 0$,

$$\begin{aligned} \frac{\partial\boldsymbol{\lambda}(t)}{\partial t} &= -\nabla_{\gamma_t}E_{sde}(t) - \nabla_{\gamma_t}\left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}(\mathbf{x})]\right\rangle_{q(\mathbf{x},t)}^{\mathrm{T}}\boldsymbol{\lambda}(t)\\ \boldsymbol{\lambda}(t_i^-) &= \boldsymbol{\lambda}(t_i) - \nabla_{\gamma_t}E_{obs}(t_i), \quad \forall t_i \in \mathcal{T}.\end{aligned}$$

[4] **Gradient step:** $\boldsymbol{\pi}_t \leftarrow \rho(\boldsymbol{\pi}_t, \nabla_{\boldsymbol{\pi}_t}\hat{\mathcal{L}}(\mathbf{g},\boldsymbol{\lambda}))$ for all $t \in [0,T]$, where

$$\nabla_{\boldsymbol{\pi}_t}\hat{\mathcal{L}}(\mathbf{g},\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\pi}_t}E_{sde}(t) + \boldsymbol{\lambda}^{\mathrm{T}}(t)\nabla_{\boldsymbol{\pi}_t}\left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}(\mathbf{x})]\right\rangle_{q(\mathbf{x},t)}.$$

[5] **Repeat:** steps [2-4] until convergence, i.e. $\nabla_{\boldsymbol{\pi}_t}\hat{\mathcal{L}}(\mathbf{g},\boldsymbol{\lambda}) = 0$.

Four main issues that need to be considered when implementing this algorithm are:

1. The choice of gradient step $\rho(\boldsymbol{\pi}_t, \nabla_{\boldsymbol{\pi}_t}\hat{\mathcal{L}}(\mathbf{g}, \boldsymbol{\lambda}))$, e.g. scaled-conjugate.
2. The choice of numerical integration method for solving the differential equations in steps [1] and [2], e.g. Euler, higher-order Runge-Kutta.
3. The choice of numerical integration method used for computing the expectations involved in steps [1] and [2], e.g. cubature methods, unscented transforms.
4. How to compute γ_0 if it is not given *a priori*.

The above algorithm relies upon $q(\mathbf{x}, t)$ being fully characterised by the moment function γ_t. If this was not the case then it would not be possible to perform any of the expectations, due to a lack of sufficient information. If $\boldsymbol{\phi}(\mathbf{x})$ is the sufficient statistic of $q(\mathbf{x}, t)$ this isn't a problem, but in the general case it is non-trivial to reconstruct a density from its moment parameter. A case where it is simple is when $q(\mathbf{x}, t)$ is Gaussian.

3.3.2 Optimal Gaussian approximation

To construct a Gaussian approximation to the posterior and subsequently a Gaussian approximation to the smoothing density, requires the drift $\mathbf{g}(\mathbf{x}, t)$ to be linear. It is in this case possible to write

$$\mathbf{g}_L(\mathbf{x}, t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{b}(t) \tag{3.21}$$

for some $\mathbf{A} : \mathbb{R}_+ \to \mathbb{R}^{d_\mathbf{x} \times d_\mathbf{x}}$ and $\mathbf{b} : \mathbb{R}_+ \to \mathbb{R}^{d_\mathbf{x}}$. Here $\mathbf{g}_L(\cdot, t)$ is parameterised by $\mathbf{A}_t \in \mathbb{R}^{d_\mathbf{x} \times d_\mathbf{x}}$ and $\mathbf{b}_t \in \mathbb{R}^{d_\mathbf{x}}$, and it is therefore possible to take gradients of the free-energy with respect to $\mathbf{A}_t$ and $\mathbf{b}_t$ at time t. Let $q_L(\mathbf{x}, t)$ denote the one-time-slice Gaussian marginal density generated using linear $\mathbf{g}_L(\mathbf{x}, t)$, let $\boldsymbol{\mu}(t)$ and $\boldsymbol{\Sigma}(t)$ denote its mean and covariance, respectively. Inserting $q_L(\mathbf{x}, t)$ and $\mathbf{g}_L(\mathbf{x}, t)$ into (3.15) leads to the set of differential moment equations

$$\frac{\partial \boldsymbol{\mu}(t)}{\partial t} = \mathbf{A}(t)\boldsymbol{\mu}(t) + \mathbf{b}(t) \tag{3.22}$$

$$\frac{\partial \boldsymbol{\Sigma}(t)}{\partial t} = \mathbf{A}(t)\boldsymbol{\Sigma}(t) + \boldsymbol{\Sigma}(t)\mathbf{A}^{\mathrm{T}}(t) + \left\langle \mathbf{D}(\mathbf{x}, t) \right\rangle_{q_L(\mathbf{x},t)}. \tag{3.23}$$

Identify $\boldsymbol{\gamma}(t)$ with $\mathrm{vec}(\boldsymbol{\mu}(t), \boldsymbol{\Sigma}(t))$ and $\boldsymbol{\lambda}(t)$ with $\mathrm{vec}(\boldsymbol{\nu}(t), \boldsymbol{\Lambda}(t))$ for some dual parameters $\boldsymbol{\nu} : \mathbb{R}_+ \to \mathbb{R}^{d_\mathbf{x}}$ and $\boldsymbol{\Lambda} : \mathbb{R}_+ \to \mathbb{R}^{d_\mathbf{x} \times d_\mathbf{x}}$. Inserting $q_L(\mathbf{x}, t)$ and $\mathbf{g}_L(\mathbf{x}, t)$ and the corresponding moment and dual parameters into equations (3.19) and (3.20) leads to the adjoint equations

$$\frac{\partial \boldsymbol{\nu}(t)}{\partial t} = -\nabla_{\boldsymbol{\mu}_t} E_{sde}(t) - \mathbf{A}^{\mathrm{T}}(t)\boldsymbol{\nu}(t) \tag{3.24}$$

$$\frac{\partial \boldsymbol{\Lambda}(t)}{\partial t} = -\nabla_{\boldsymbol{\Sigma}_t} E_{sde}(t) - \mathbf{A}^{\mathrm{T}}(t)\boldsymbol{\Lambda}(t) - \boldsymbol{\Lambda}(t)\mathbf{A}^{\mathrm{T}}(t) \tag{3.25}$$

and jump conditions at the observations times $t_i \in [0, T]$, given by

$$\begin{aligned}
\boldsymbol{\nu}(t_i) &= \boldsymbol{\nu}(t_i^+) - \nabla_{\boldsymbol{\mu}_t} E_{obs}(t_i) && (3.26)\\
\boldsymbol{\Lambda}(t_i) &= \boldsymbol{\Lambda}(t_i^+) - \nabla_{\boldsymbol{\Sigma}_t} E_{obs}(t_i). && (3.27)
\end{aligned}$$

Now identifying $\boldsymbol{\pi}(t)$ with $\mathrm{vec}(\mathbf{A}(t), \mathbf{b}(t))$, and inserting $q_L(\mathbf{x}, t)$ and $\mathbf{g}_L(\mathbf{x}, t)$ and the corresponding moment and dual parameters into equation (3.17) leads to the gradient equations

$$\begin{aligned}
\nabla_{\mathbf{A}_t} \hat{\mathcal{L}}(\mathbf{g}, \boldsymbol{\lambda}) &= \nabla_{\mathbf{A}_t} E_{sde}(t) + \boldsymbol{\Sigma}(t)\boldsymbol{\Lambda}^{\mathrm{T}}(t) + \boldsymbol{\Lambda}(t)\boldsymbol{\Sigma}(t) + \boldsymbol{\nu}(t)\boldsymbol{\mu}^{\mathrm{T}}(t) && (3.28)\\
\nabla_{\mathbf{b}_t} \hat{\mathcal{L}}(\mathbf{g}, \boldsymbol{\lambda}) &= \nabla_{\mathbf{b}_t} E_{sde}(t) + \boldsymbol{\nu}(t). && (3.29)
\end{aligned}$$

Equations (3.24),(3.25),(3.26),(3.27),(3.28) and (3.29), completely determine the variational Gaussian approximate smoothing algorithm derived in Archambeau et al. (2007b). Thus the variational Gaussian-process smoothing algorithm is a natural specialisation of the more general smoothing algorithm described in the first part of this section.

3.4 Projection filter

Consider the partial-differential equation in (3.13). For any twice differentiable map $\boldsymbol{\phi}(\mathbf{x})$, equation (3.13) describes the time-evolution of expectations of $\boldsymbol{\phi}(\mathbf{x})$ under $q(\mathbf{x}, t)$, where $q(\mathbf{x}, t)$ is the one-time marginal density generated by the SDE in equation (2.21) with drift function $\mathbf{g}(\mathbf{x}, t)$. Now assume that $\mathbf{g}(\mathbf{x}, t)$ is of a form that the marginal density $q(\mathbf{x}, t)$ is intractable for some, or all, of $t \in [0, T]$. Under certain conditions on $\mathbf{g}(\mathbf{x}, t)$ and $\mathbf{D}(\mathbf{x}, t)$, the marginal density $q(\cdot, t)$ can be constrained to lie in a well-behaved function space, such as $L^1(\mathbb{R}^{d_\mathbf{x}}; \mathbb{R})$. But given the infinite degrees of freedom in such a space, it may not be possible to capture the properties of $q(\cdot, t)$ in a finite description length. Given that $q(\cdot, t)$ is required for all $t \in [0, T]$, the problem of representing the marginal densities for general continuous-time systems quickly becomes infeasible, even under simplifying conditions.

The crux of the *Projection filter* (Brigo et al., 1999) is to obtain a finite-dimensional projection of the marginal density $q(\cdot, t)$ onto a parameterised manifold lying in an ambient space, such as $L^1(\mathbb{R}^{d_\mathbf{x}}; \mathbb{R})$. In this section of the thesis it is assumed that any observations have already been assimilated into the drift $\mathbf{g}(\mathbf{x}, t)$. It is the mathematical properties of the projection that are of importance. While the machinery below might seem a little overzealous, it lays out a principled geometric derivation of the projection map and its application to the Fokker-Planck equation. An identical result can be achieved simply by replacing the expectations in (3.13) with expectations under a simplified model with sufficient statistic $\boldsymbol{\phi}(\mathbf{x})$. While this approach leads to an (observation-free) version of the commonly used *Assumed Density Filter* (ADF) of Kushner (1967), this algorithm is considered heuristic, even by its originator many years on (Kushner, 2008). Therefore the reason for presenting the following formulation is, in some way, to expose the credibility of the projection map before applying it to the variational smoothing problem. If the derivation is not to the readers tastes, then a novel derivation, requiring less technical machinery, is provided in section 3.5.

3.4.1 Projection map

The current section outlines the correct method for projecting probability densities onto a finite-dimensional manifold, while keeping the amount of formal differential-geometry to a minimum. For more details see Brigo et al. (1995) and Brigo et al. (1999), or Berefelt et al. (2003) for a nice overview.

For an open subset $\boldsymbol{\Theta} \subset \mathbb{R}^{d_{\boldsymbol{\theta}}}$, let $S := \{p(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ denote a set of strictly-positive Lebesgue measurable probability density functions in $L^1(\mathbb{R}^{d_{\mathbf{x}}}; \mathbb{R})$. Let θ_i denote the i^{th} component of $\boldsymbol{\theta}$. Assuming $p(\cdot, \boldsymbol{\theta})$ to be smooth in $\boldsymbol{\theta}$, the Fréchet derivative $\frac{\partial p(\cdot,\boldsymbol{\theta})}{\partial \theta_i}$ maps vectors in $\mathbb{R}^{d_{\boldsymbol{\theta}}}$ to vectors in $L^1(\mathbb{R}^{d_{\mathbf{x}}}; \mathbb{R})$. If the set of vectors $\{\frac{\partial p(\cdot,\boldsymbol{\theta})}{\partial \theta_i}\}_i$ are linearly independent for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, then S is a *submanifold* of $L^1(\mathbb{R}^{d_{\mathbf{x}}}; \mathbb{R})$, with a tangent space T_pS at each point $p \in S$ spanned by the coordinates $\{\frac{\partial}{\partial \theta_i}\}_i$. To turn S into a Riemannian manifold, a metric tensor $\mathbf{G}_p(\cdot, \cdot)$ is defined on the tangent space T_pS such that

$$\mathbf{G}_p(\frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_i}, \frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_j}) = \int \frac{\partial p(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial p(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_j} \frac{d\mathbf{x}}{p(\mathbf{x}, \boldsymbol{\theta})}. \tag{3.30}$$

This integral is equivalent to the $(i, j)^{th}$ entry of the well known *Fisher information* matrix of statistics. Define the matrix $\mathbf{G}(\boldsymbol{\theta})$ with $(i, j)^{th}$ entry $\mathbf{G}_{ij}(\boldsymbol{\theta}) := \mathbf{G}_p(\frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_i}, \frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_j})$ and let $\mathbf{G}^{ij}(\boldsymbol{\theta})$ denote the $(i, j)^{th}$ entry of the inverse $\mathbf{G}^{-1}(\boldsymbol{\theta})$. Even though the tensor $\mathbf{G}_p(u, v)$ is defined only for vectors $u, v \in T_pS$, the product

$$\bar{\mathbf{G}}_p(u, z) = \int u(\mathbf{x}) z(\mathbf{x}) \frac{d\mathbf{x}}{p(\mathbf{x}, \boldsymbol{\theta})} \tag{3.31}$$

may exist for pairs of vectors $u, z \in L^1$ such that $u \in T_pS$ and z does not. Let $\bar{\mathbf{G}}_p$ denote the tensor $\mathbf{G}_p$ with the second argument extended to all $z \in L^1$ and its first argument still in T_pS. The extended domain D_p is a linear subspace of L^1 characterised by

$$D_p := \{z \in L^1 | \frac{z}{p_{\boldsymbol{\theta}}} \frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_i} \in L^1\}. \tag{3.32}$$

Then for any $p \in S$ and any $z \in D_p$ it is possible to define a projection $\Pi_{\boldsymbol{\theta}} : D_p \to T_pS$ such that

$$\Pi_{\boldsymbol{\theta}}(z) := \sum_{i,j=1}^{d_{\boldsymbol{\theta}}} \mathbf{G}^{ij}(\boldsymbol{\theta}) \bar{\mathbf{G}}_p(\frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_j}, z) \frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_i}. \tag{3.33}$$

This is the map used to define the projection filter[2].

[2] If equation (3.33) seems a little cryptic, then the reader should bare in mind a Euclidean version where the ambient space $\mathbb{R}^n$ is spanned by the basis vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ and the subspace $V = \text{span}\{\mathbf{e}_i\}_{i \in \mathbf{i}}$ defines a

3.4.2 Projected Fokker-Planck equation

In this section the projected Fokker-Planck equation is derived for a parametric family of probability densities. The method is first discussed in the general setting and then, in the sequel, specialised to the case of a minimal exponential family.

Let $S := \{p(\cdot,\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be a family of probability densities, for open subset $\boldsymbol{\Theta} \subset \mathbb{R}^{d_{\boldsymbol{\theta}}}$. Let $\boldsymbol{\theta} \in \Theta$ be a function of time $\boldsymbol{\theta}(t)$, such that $\boldsymbol{\theta}(t) \in \Theta$ for all $t \in [0,T]$. Using the (Stratonovich) chain rule, it holds that

$$\frac{\partial p_{\boldsymbol{\theta}(t)}}{\partial t} = \sum_{i=1}^{d_{\boldsymbol{\theta}}} \frac{\partial \theta_i(t)}{\partial t} \frac{\partial p_{\boldsymbol{\theta}(t)}}{\partial \theta_i(t)}. \tag{3.34}$$

If we impose the condition

$$\frac{\overrightarrow{\mathcal{K}}_{\mathbf{g}}[p_{\boldsymbol{\theta}}]}{p_{\boldsymbol{\theta}}} \frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_i} \in L^1, \quad i = 1, \ldots, d_{\boldsymbol{\theta}}, \tag{3.35}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $t \in [0,T]$, then the projection $\Pi_{\boldsymbol{\theta}}$ can be applied directly to the function $\overrightarrow{\mathcal{K}}_{\mathbf{g}}[p_{\boldsymbol{\theta}}]$. More precisely,

$$\begin{aligned}
\Pi_{\boldsymbol{\theta}(t)}\Big(\overrightarrow{\mathcal{K}}_{\mathbf{g}}[p_{\boldsymbol{\theta}(t)}]\Big) &= \sum_{i,j=1}^{d_{\boldsymbol{\theta}}} \mathbf{G}^{ij}(\boldsymbol{\theta}(t))\bar{\mathbf{G}}_p(\frac{\partial p_{\boldsymbol{\theta}(t)}}{\partial \theta_j(t)}, \overrightarrow{\mathcal{K}}_{\mathbf{g}}[p_{\boldsymbol{\theta}(t)}])\frac{\partial p_{\boldsymbol{\theta}(t)}}{\partial \theta_i(t)} && (3.36)\\
&= \sum_{i,j=1}^{d_{\boldsymbol{\theta}}} \mathbf{G}^{ij}(\boldsymbol{\theta}(t)) \int \frac{\partial p_{\boldsymbol{\theta}(t)}}{\partial \theta_j(t)} \overrightarrow{\mathcal{K}}_{\mathbf{g}}[p_{\boldsymbol{\theta}(t)}] \frac{d\mathbf{x}}{p_{\boldsymbol{\theta}(t)}} \frac{\partial p_{\boldsymbol{\theta}(t)}}{\partial \theta_i(t)} && (3.37)
\end{aligned}$$

where the dependency of $p_{\boldsymbol{\theta}(t)}$ on $\mathbf{x}$ is implicit. Observing equations (3.34) and (3.37), we see that both are built on the set of coordinate vectors $\{\frac{\partial p_{\boldsymbol{\theta}(t)}}{\partial \theta_i(t)}\}_i$. Therefore changes in one can easily be mapped to changes in the other simply by matching the coefficients in the local coordinates. Indeed, let us define the *projected Fokker-Planck* equation

$$\frac{\partial p_{\boldsymbol{\theta}(t)}}{\partial t} = \Pi_{\boldsymbol{\theta}(t)}\Big(\overrightarrow{\mathcal{K}}_{\mathbf{g}}[p_{\boldsymbol{\theta}(t)}]\Big). \tag{3.38}$$

Equating coefficients in the basis $\{\frac{\partial p_{\boldsymbol{\theta}(t)}}{\partial \theta_i(t)}\}_i$, leads to an evolution equation in the parameter space $\boldsymbol{\Theta}$, given by

$$\frac{\partial \boldsymbol{\theta}(t)}{\partial t} = \mathbf{G}^{-1}(\boldsymbol{\theta}) \int \frac{\partial p_{\boldsymbol{\theta}(t)}}{\partial \boldsymbol{\theta}(t)} \overrightarrow{\mathcal{K}}_{\mathbf{g}}[p_{\boldsymbol{\theta}(t)}] \frac{d\mathbf{x}}{p_{\boldsymbol{\theta}(t)}} \tag{3.39}$$

where $\frac{\partial \boldsymbol{\theta}(t)}{\partial t}$ and $\frac{\partial p(\mathbf{x},\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}(t)} \equiv \nabla_{\boldsymbol{\theta}(t)} p(\mathbf{x}, \boldsymbol{\theta}(t))$ are vectors in $\mathbb{R}^{d_{\boldsymbol{\theta}}}$ and $\overrightarrow{\mathcal{K}}_{\mathbf{g}}[p_{\boldsymbol{\theta}(t)}]$ is a function in L^1.

submanifold for any $\mathbf{i} \subseteq \mathbb{R}^n$. In this case $\mathbf{G}_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j \rangle = \delta_{ij}$ and a projection $\Pi : \mathbb{R}^n \to V$ can be defined simply by $\Pi(\mathbf{z}) = \sum_{i \in \mathbf{i}} \langle \mathbf{e}_i, \mathbf{z} \rangle \mathbf{e}_i$.

3.4.3 Exponential-family Fokker-Planck projection

Now let us specialise the above formulation to exponential families. Let $\boldsymbol{\phi} : \mathbb{R}^{d_{\mathbf{x}}} \to \mathbb{R}^{d_{\boldsymbol{\theta}}}$ be a twice differentiable vector of sufficient statistics and define

$$p(\mathbf{x}, \boldsymbol{\theta}) = \exp\Big(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})\Big), \tag{3.40}$$

with log partition $A(\boldsymbol{\theta})$ and domain $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}} | A(\boldsymbol{\theta}) < \infty\}$. To interpret $S := \{p(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ as a $d_{\boldsymbol{\theta}}$-dimensional manifold, the components of $\boldsymbol{\phi}$ must be linearly independent. Let $\boldsymbol{\gamma}(t) = \langle \phi(\mathbf{x}) \rangle_{p(\mathbf{x}, \boldsymbol{\theta}(t))}$ denote the vector of mean parameters of $p(\mathbf{x}, \boldsymbol{\theta}(t))$. Assuming the condition in equation (3.35) holds, we have (see Brigo et al. (1999))

$$\int \frac{\partial p_{\boldsymbol{\theta}(t)}}{\partial \boldsymbol{\theta}(t)} \overrightarrow{\mathcal{K}}_{\mathbf{g}}[p_{\boldsymbol{\theta}(t)}] \frac{d\mathbf{x}}{p_{\boldsymbol{\theta}(t)}} \quad = \quad \big\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}] \big\rangle_{p_{\boldsymbol{\theta}(t)}}. \tag{3.41}$$

Inserting equation (3.41) into (3.39), and noting the relation $d\boldsymbol{\gamma}(t) = \mathbf{G}(\boldsymbol{\theta}(t))d\boldsymbol{\theta}(t)$, the projection filter can be defined in moment form. In the following, let $S := \{p(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ denote an $d_{\boldsymbol{\theta}}$-dimensional exponential family manifold and assume condition (3.35) holds.

Definition 3.1 (Projection filter)**.** *The (observation-free) Projection filter for the SDE in* (2.21) *with drift function* $\mathbf{g}(\mathbf{x}, t)$*, defines the set of marginal densities* $\{p(\mathbf{x}, \boldsymbol{\theta}(t)) \equiv p(\mathbf{x}, \boldsymbol{\gamma}(t)) \in S, t \in [0, T]\}$ *such that* $\boldsymbol{\gamma}(t)$ *and* $\boldsymbol{\theta}(t)$ *satisfy the differential equations*

$$\frac{\partial \boldsymbol{\gamma}(t)}{\partial t} = \mathbf{G}(\boldsymbol{\theta}(t)) \frac{\partial \boldsymbol{\theta}(t)}{\partial t} = \big\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}] \big\rangle_{p_{\boldsymbol{\theta}(t)}}. \tag{3.42}$$

It is important to identify how equation (3.42) differs from equation (3.15). Equation (3.15) follows directly from the the Fokker-Planck equation (2.24), and the marginal densities in equation (3.15) are those directly generated by the SDE in equation (2.21) with drift function $\mathbf{g}(\mathbf{x}, t)$. In contrast, equation (3.42) is derived from the projection of the Fokker-Planck equation onto the submanifold S. The Fokker-Planck equation defines a vector field on the space of probabilities. Equation (3.42) projects this vector field onto the tangent space at each point of S and lets $\boldsymbol{\theta}(t)$ evolve under a map of the projected field back into $\boldsymbol{\Theta}$. The marginal densities in (3.42) are the result of this projected action of inference and remain in S for all $t \in [0, T]$. While this ensures that their evolution is guided by the drift $\mathbf{g}(\mathbf{x}, t)$, it also ensures that they remain tractable, inasmuch as having a finite dimensional representation. In the sequel, these tractable marginals will provide a feasible domain for variational approximate inference, and the projected moment equations in (3.42) will provide suitable moment constraints on such a domain. For equation (3.42) to be complete, initial conditions $\boldsymbol{\gamma}(0) = \boldsymbol{\gamma}_0$ and $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$ are required. Assume an initial distribution $p(\mathbf{x}_0) = p(\mathbf{x}, 0)$ is provided with the SDE in equation (2.21). As recommended in Brigo et al. (1999), it makes sense to define

$$\boldsymbol{\gamma}_0 = \big\langle \boldsymbol{\phi}(\mathbf{x}) \big\rangle_{p(\mathbf{x},0)}. \tag{3.43}$$

This is equivalent to minimising the relative entropy between $p(\mathbf{x}, 0)$ and $p(\mathbf{x}, \boldsymbol{\theta}_0)$ over $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$. This ensures that the projected density $p(\mathbf{x}, \boldsymbol{\theta}(t))$ begins in S and remains in S when evolving according to equation (3.42). The initial condition $\boldsymbol{\theta}_0$ can be obtained from γ_0 through the backward-mapping in equation (A.8).

3.4.4 Original projection-filter

The projection-filter of (Brigo et al., 1999) in its original formulation was not designed just to trace the moments of an intractable marginal density. It was also designed to assimilate data and to approximate the filtering density $p(\mathbf{x}_t|\mathbf{Y}_{\leq t})$. Let $\boldsymbol{\gamma}(t)$ denote the moment parameter of a parameterised probability density $p_{\boldsymbol{\gamma}(t)}(\mathbf{x})$ with corresponding sufficient statistic $\boldsymbol{\phi}(\mathbf{x})$. Let t^- denote a time infinitesimally close to time t. In between observation times the parameter $\boldsymbol{\gamma}(t)$ evolves according to equation (3.42) using a *prior* drift $\mathbf{f}(\mathbf{x}, t)$, chosen a priori. At an observation time t_i, $\boldsymbol{\gamma}(t)$ is updated according to the equation

$$\boldsymbol{\gamma}(t_i) = \frac{\int \boldsymbol{\phi}(\mathbf{x}) p(\mathbf{y}_i|\mathbf{x}) p(\mathbf{x}, \boldsymbol{\gamma}_{t_i^-}) d\mathbf{x}}{\int p(\mathbf{y}_i|\mathbf{x}) p(\mathbf{x}, \boldsymbol{\gamma}_{t_i^-}) d\mathbf{x}} \tag{3.44}$$

which follows naturally from Bayes' rule. The simplicity of the projection filter makes it a very fast method for assimilating data. Importantly, it is defined in continuous-time and the differential equations in the propagation step can be solved using any appropriate numerical integration method. This projection filter takes a more common form (Maybeck, 1979) when specialised to the Gaussian setting.

3.4.4.1 Gaussian projection-filter

Let $\boldsymbol{\mu}(t)$ and $\boldsymbol{\Sigma}(t)$ denote the mean and covariance parameters of the one-slice marginal Gaussian density $q(\mathbf{x}, t)$, and identify $\boldsymbol{\gamma}(t)$ with $\mathrm{vec}(\boldsymbol{\mu}(t), \boldsymbol{\Sigma}(t))$. The propagation step of the projection-filter then leads directly to mean and covariance evolution equations

$$\begin{aligned} \frac{\partial \boldsymbol{\mu}(t)}{\partial t} &= \left\langle \mathbf{f}(\mathbf{x}, t) \right\rangle_{p(\mathbf{x}, \boldsymbol{\gamma}_t)} && (3.45) \\ \frac{\partial \boldsymbol{\Sigma}(t)}{\partial t} &= \left\langle \mathbf{f}(\mathbf{x}, t)(\mathbf{x} - \boldsymbol{\mu}_t)^{\mathrm{T}} + (\mathbf{x} - \boldsymbol{\mu}_t)\mathbf{f}^{\mathrm{T}}(\mathbf{x}, t) \right\rangle_{p(\mathbf{x}, \boldsymbol{\gamma}_t)} + \left\langle \mathbf{D}^2(\mathbf{x}, t) \right\rangle_{p(\mathbf{x}, \boldsymbol{\gamma}_t)}. && (3.46) \end{aligned}$$

The observation model corresponding to the free-energy formulation of chapter (3) is given in state-space form by

$$\mathbf{y}_i = \mathbf{h}(\mathbf{x}_{t_i}, t_i) + \mathbf{R}(\mathbf{x}_{t_i}, t_i)\boldsymbol{\eta}_i, \qquad t_i \in \mathcal{T}, \tag{3.47}$$

where $\mathbf{h} : \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}_+ \to \mathbb{R}^{d_\mathbf{y}}$ denotes a nonlinear observation mapping, $\mathbf{R} : \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}_+ \to \mathbb{R}^{d_\mathbf{y} \times d_\mathbf{y}}$ denotes a positive-definite noise matrix, and each $\boldsymbol{\eta}_i$ is Gaussian distributed with zero mean and unit variance. It follows that $p(\mathbf{y}_i|\mathbf{x}_{t_i})$ is conditionally Gaussian, and the update step of the projection-filter can be solved exactly using the rules for conditioning Gaussians, to give jump conditions

$$\begin{aligned}\boldsymbol{\mu}(t_i) &= \boldsymbol{\mu}(t_i^-) + \mathbf{C}_i\mathbf{S}_i^{-1}(\mathbf{y}_i - \bar{\mathbf{y}}_i) &(3.48)\\ \boldsymbol{\Sigma}(t_i) &= \boldsymbol{\Sigma}(t_i^-) - \mathbf{C}_i\mathbf{S}_i\mathbf{C}_i^{\mathrm{T}} &(3.49)\end{aligned}$$

where

$$\begin{aligned}\bar{\mathbf{y}}_i &= \big\langle \mathbf{h}(\mathbf{x}, t_i)\big\rangle_{p(\mathbf{x},t_i^-)} &(3.50)\\ \mathbf{S}_i &= \Big\langle \big(\mathbf{h}(\mathbf{x}, t_i) - \bar{\mathbf{y}}_i\big)\big(\mathbf{h}(\mathbf{x}, t_i) - \bar{\mathbf{y}}_i\big)^{\mathrm{T}}\Big\rangle_{q(\mathbf{x},t_i^-)} + \big\langle \mathbf{R}^2(\mathbf{x}, t)\big\rangle_{q(\mathbf{x},t_i^-)} &(3.51)\\ \mathbf{C}_i &= \Big\langle \big(\mathbf{x} - \boldsymbol{\mu}(t_i^-)\big)\big(\mathbf{h}(\mathbf{x}, t_i) - \bar{\mathbf{y}}_i\big)^{\mathrm{T}}\Big\rangle_{q(\mathbf{x},t_i^-)}. &(3.52)\end{aligned}$$

These latter equations simply propagate the current density $q(\mathbf{x}, t_i^-)$ through the nonlinear observation model. They contribute to a joint distribution, predicting where $\mathbf{y}$ is expected to be based on $q(\mathbf{x}, t_i^-)$, and then equations (3.48) and (3.49) simply condition the joint distribution on the realisation of $\mathbf{y}$. To implement the GP filter in the general setting, the expectations in the above equations need to be approximated. When $\mathbf{f}(\mathbf{x}, t)$ is differentiable, the expectations in equation (3.46) can be simplified using the identity (appendix A.3.3)

$$\big\langle \mathbf{f}(\mathbf{x}, t)(\mathbf{x} - \boldsymbol{\mu}_t)^{\mathrm{T}}\big\rangle = \big\langle \nabla_{\mathbf{x}}\mathbf{f}(\mathbf{x}, t)\big\rangle\boldsymbol{\Sigma}(t). \qquad (3.53)$$

3.5 Assumed density filter

The derivation of the projection filter requires a large amount of technical differential-geometry. This section shows how the projection filter can also be defined as the continuous time limit of a simple forward algorithm applied to a linear discretisation of (2.21). First the proof of the Fokker-Planck equation is given as of Jazwinski (1970) and Risken (1996). The proof helped guide the novel derivation of the projection filter and is presented here for completeness, and to show how only a simple projection step needs to be incorporated into the forward algorithm to obtain the projection filter of the previous section.

3.5.1 Fokker-Planck revisited

Let $p(\mathbf{x}, t)$ denote the marginal density at time t over paths $\mathbf{x}_t$ of (2.21). After some time step δt, the marginal density $p(\mathbf{x}, t + \delta t)$ is given by the Chapman-Kolmogorov equation

$$p(\mathbf{x}, t + \delta t) = \int p(\mathbf{x}, t + \delta t|\mathbf{x}', t)p(\mathbf{x}', t)d\mathbf{x}' \qquad (3.54)$$

where the transition density $p(\mathbf{x}, t|\mathbf{x}', t')$ may or may not have closed form. Alternatively, using a first-order Taylor's expansion of $p(\mathbf{x}, t + \delta t)$ in terms of t, $p(\mathbf{x}, t + \delta t)$ can be written

$$p(\mathbf{x}, t + \delta t) = p(\mathbf{x}, t) + \frac{\partial p(\mathbf{x}, t)}{\partial t}\delta t + o(\delta t). \tag{3.55}$$

For any function $u(\mathbf{x})$, such that following integral exists, equation (3.55) leads to the relation

$$\left\langle \delta t \frac{\partial p(\mathbf{x}, t)}{\partial t} + o(\delta t) \right\rangle_{u(\mathbf{x})} = \big\langle p(\mathbf{x}, t + \delta t) - p(\mathbf{x}, t) \big\rangle_{u(\mathbf{x})}. \tag{3.56}$$

Using equation (3.54), and using Fubini's theorem to interchange the integrals, the right-hand-side of (3.56) can be written

$$\big\langle p(\mathbf{x}, t + \delta t) - p(\mathbf{x}, t) \big\rangle_{u(\mathbf{x})} \;\; = \;\; \left\langle \big\langle u(\mathbf{x}') \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)} - u(\mathbf{x}) \right\rangle_{p(\mathbf{x}, t)} \tag{3.57}$$

where the *dummy* variables $\mathbf{x}'$ and $\mathbf{x}$ have been exchanged in the inner expectation. Now assume that $u(\mathbf{x})$ is twice differentiable and vanishes at infinity. The first property allows us to use a Taylor expansion of $u(\mathbf{x}')$ around $\mathbf{x}$, up to second-order terms. This allows the inner expectation on the r.h.s. of (3.57) to be written

$$\begin{aligned} \big\langle u(\mathbf{x}') \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)} &= u(\mathbf{x}) + \nabla_{\mathbf{x}}^{\mathrm{T}} u(\mathbf{x}) \big\langle \delta \mathbf{x} \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)} \\ &\quad + \frac{1}{2}\mathrm{tr}\Big(\nabla_{\mathbf{x}} \nabla_{\mathbf{x}}^{\mathrm{T}} u(\mathbf{x}) \big\langle \delta\mathbf{x}\delta\mathbf{x}^{\mathrm{T}} \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)} \Big) + \big\langle o(\delta\mathbf{x}\delta\mathbf{x}^{\mathrm{T}}) \big\rangle \end{aligned} \tag{3.58}$$

where $\delta\mathbf{x} := \mathbf{x}' - \mathbf{x}$ denotes a realisation of $\mathbf{x}_{t+\delta t} - \mathbf{x}_t$. Now assume that δt is small enough that (2.21) can be replaced with a first-order Euler approximation

$$\mathbf{x}(t + \delta t) = \mathbf{x}(t) + \delta t \mathbf{g}(\mathbf{x}(t), t) + \sqrt{\delta t \mathbf{D}(\mathbf{x}(t), t)} \boldsymbol{\xi}_t, \tag{3.59}$$

where $\boldsymbol{\xi}_t = \int_0^{\delta t} d\mathbf{w}_t$ represents the system noise in the discretised model. From the Gaussian properties of $\boldsymbol{\xi}_t$, it holds that the expectations in equation (3.58) equate to

$$\begin{aligned} \big\langle \delta\mathbf{x} \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)} &= \delta t \mathbf{g}(\mathbf{x}(t), t) && (3.60) \\ \big\langle \delta\mathbf{x}\delta\mathbf{x}^{\mathrm{T}} \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)} &= \delta t \mathbf{D}(\mathbf{x}(t), t) + o(\delta t) && (3.61) \\ \big\langle o(\delta\mathbf{x}\delta\mathbf{x}^{\mathrm{T}}) \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)} &= o(\delta t). && (3.62) \end{aligned}$$

Inserting these into equation (3.58) leads to

$$\begin{aligned} \big\langle u(\mathbf{x}') \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)} &= u(\mathbf{x}) + \delta t \mathbf{g}^{\mathrm{T}}(\mathbf{x}(t), t) \nabla_{\mathbf{x}} u(\mathbf{x}) \\ &\quad + \delta t \frac{1}{2}\mathrm{tr}\Big(\mathbf{D}(\mathbf{x}(t), t) \nabla_{\mathbf{x}} \nabla_{\mathbf{x}}^{\mathrm{T}} u(\mathbf{x}) \Big) + o(\delta t) && (3.63) \\ &= u(\mathbf{x}) + \delta t \overleftarrow{\mathcal{K}}_{\mathbf{g}}[u(\mathbf{x})] + o(\delta t). && (3.64) \end{aligned}$$

Which, substituting back into (3.56), leads to the relation

$$\left\langle \delta t \frac{\partial p(\mathbf{x},t)}{\partial t} + o(\delta t) \right\rangle_{u(\mathbf{x})} = \left\langle \delta t \overset{\leftarrow}{\mathcal{K}}_{\mathbf{g}}[u(\mathbf{x})] + o(\delta t) \right\rangle_{p(\mathbf{x},t)}. \tag{3.65}$$

Dividing through by δt, taking the limit $\delta t \to 0$ gives

$$\left\langle \frac{\partial p(\mathbf{x},t)}{\partial t} \right\rangle_{u(\mathbf{x})} = \left\langle \overset{\leftarrow}{\mathcal{K}}_{\mathbf{g}}[u(\mathbf{x})] \right\rangle_{p(\mathbf{x},t)}. \tag{3.66}$$

Performing integration by parts on the left hand side (using the dual-adjoint nature of $\overset{\leftarrow}{\mathcal{K}}_{\mathbf{g}}$ and $\overset{\rightarrow}{\mathcal{K}}_{\mathbf{g}}$, and the fact that $u(\mathbf{x})$ and $p(\mathbf{x},t)$ both vanish at infinity) gives

$$\left\langle \left(\partial_t - \overset{\rightarrow}{\mathcal{K}}_{\mathbf{g}}\right)[p(\mathbf{x},t)] \right\rangle_{u(\mathbf{x})} = 0. \tag{3.67}$$

Since $u(\mathbf{x})$ was arbitrary this proves (2.24).

3.5.2 Assumed density projection

In this section the projection filter of section (3.4) is derived from a different perspective. It is motivated by the Bayesian approach to online learning of Opper (1998) and the more general framework of Expectation Propagation Minka (2001). These algorithms were traditionally designed for discrete-time variables, but by taking the limiting case, as in the exact inference problem of the previous section, one can obtain analogous (continuous-time) approximate inference solutions. Importantly, the proof does not require the assumption of an canonical exponential family.

Refer back to the Chapman-Kolmogorov equation given in (3.54). If equation (3.54) is applied repeatedly, for general $p(\mathbf{x}, t+\delta t|\mathbf{x}', t)$, the marginal densities $p(\mathbf{x},t)$ can quickly become intractable. This means the marginals have either no closed form or one that is computationally inefficient to store. The idea of Assumed Density Filtering (ADF) is to assume that $p(\mathbf{x}',t)$ belongs to some tractable family of densities, and to project $p(\mathbf{x}, t+\delta t)$ back onto this family after the propagation step in (3.54). This leads to an approximate inference algorithm that iterates between inference and projection, slowly moving along the discretised sequence of times. A small time step is taken and the result is projected back onto the tractable family. In the limit this results in an evolution equation for the vector of moment parameters analogous to the projection filter of equation (3.42). More formally, Let $\boldsymbol{\phi}(\mathbf{x})$ denote the sufficient statistic of a family $\{q(\mathbf{x},\boldsymbol{\gamma}(t)), \boldsymbol{\gamma}(t) \in \mathcal{M}\}_{t\in[0,T]}$ with mean parameter $\boldsymbol{\gamma}(\cdot)$. Assume $q(\mathbf{x},\boldsymbol{\gamma}(t))$ is known. The ADF algorithm first performs an *update* step:

$$\hat{p}(\mathbf{x}, t+\delta t) = \int p(\mathbf{x}, t+\delta t|\mathbf{x}', t) q(\mathbf{x}', \boldsymbol{\gamma}(t)) d\mathbf{x}'. \tag{3.68}$$

It then performs a *projection* step:

$$\boldsymbol{\gamma}(t+\delta t) = \int \boldsymbol{\phi}(\mathbf{x})\hat{p}(\mathbf{x}, t+\delta t)d\mathbf{x}. \tag{3.69}$$

As mentioned previously, the projection step equates to minimising the relative entropy between $\hat{p}(\mathbf{x}, t+\delta t)$ and $q(\mathbf{x}, \boldsymbol{\gamma}(t+\delta t))$. The algorithm is initialised with $\boldsymbol{\gamma}(0) = \langle\boldsymbol{\phi}(\mathbf{x})\rangle_{p(\mathbf{x},0)}$ and iterates through the update and projection steps to obtain marginal densities for all times on the discretisation. If continuous-time solutions are required the discretisation step δt can be made arbitrarily small. In this case, equation (2.21) can be replaced with the first-order Euler approximation given in (3.59). For simplicity let $\phi(\mathbf{x})$ be real valued, the following argument can be applied to each component of $\boldsymbol{\phi}(\mathbf{x})$ individually. Assume $\phi(\mathbf{x})$ is twice differentiable. Let us replace the arbitrary $u(\mathbf{x})$ in the proof of the previous section with $\phi(\mathbf{x})$. Combing the update and projection steps into one equation and using Fubini's theorem to exchange the integrals, we can write

$$\begin{aligned}
\boldsymbol{\gamma}(t+\delta t) &= \int \boldsymbol{\phi}(\mathbf{x}) \int p(\mathbf{x}, t+\delta t|\mathbf{x}', t)q(\mathbf{x}', \boldsymbol{\gamma}(t))d\mathbf{x}'d\mathbf{x} && (3.70)\\
&= \Big\langle \big\langle \boldsymbol{\phi}(\mathbf{x}') \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)} \Big\rangle_{q(\mathbf{x}, \boldsymbol{\gamma}(t))}, && (3.71)
\end{aligned}$$

where the dummy variables $\mathbf{x}'$ and $\mathbf{x}$ have been exchanged. Applying a second-order Taylor expansion to $\phi(\mathbf{x})$, the inner expectation in (3.71) equates to

$$\begin{aligned}
\big\langle \phi(\mathbf{x}') \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)} &= \phi(\mathbf{x}) + \nabla_{\mathbf{x}}^{\mathrm{T}}\phi(\mathbf{x})\big\langle \delta\mathbf{x} \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)}\\
&\quad + \frac{1}{2}\mathrm{tr}\Big(\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}^{\mathrm{T}}\phi(\mathbf{x})\big\langle \delta\mathbf{x}\delta\mathbf{x}^{\mathrm{T}} \big\rangle_{p(\mathbf{x}', t+\delta t|\mathbf{x}, t)}\Big) + \big\langle o(\delta\mathbf{x}\delta\mathbf{x}^{\mathrm{T}})\big\rangle (3.72)
\end{aligned}$$

where $\delta\mathbf{x} := \mathbf{x}' - \mathbf{x}$ denotes a realisation of $\mathbf{x}_{t+\delta t} - \mathbf{x}_t$. Inserting equation (3.72) into equation (3.71), and using the identities in equations (3.60), (3.61), and (3.62), leads to the relation

$$\begin{aligned}
\boldsymbol{\gamma}(t+\delta t) &= \Big\langle \phi(\mathbf{x}) + \delta t \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\phi(\mathbf{x})] + o(\delta t) \Big\rangle_{q(\mathbf{x}, \boldsymbol{\gamma}(t))} && (3.73)\\
&= \boldsymbol{\gamma}(t) + \Big\langle \delta t \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\phi(\mathbf{x})] \Big\rangle_{q(\mathbf{x}, \boldsymbol{\gamma}(t))} + o(\delta t). && (3.74)
\end{aligned}$$

Rearranging, we obtain

$$\frac{\boldsymbol{\gamma}(t+\delta t) - \boldsymbol{\gamma}(t)}{\delta t} = \Big\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\phi(\mathbf{x})] \Big\rangle_{q(\mathbf{x}, \boldsymbol{\gamma}(t))} + \frac{o(\delta t)}{\delta t} \tag{3.75}$$

and in the limit $\delta t \to 0$, this equation is equivalent to a one-dimensional version of the evolution equation in (3.42). Given the above argument can be applied to each component of $\boldsymbol{\phi}(\mathbf{x})$, the proof is easily extended to obtain the full multivariate version of (3.42). This algorithm is also initialised with $\boldsymbol{\gamma}_0 = \big\langle \boldsymbol{\phi}(\mathbf{x}) \big\rangle_{p(\mathbf{x},0)}$. It was not required to assume that the components of $\boldsymbol{\phi}(\mathbf{x})$

vanish at infinity because integration-by-parts was not used.

3.6 Variational projection-smoother

In the variational formulation of chapter 3, the drift function $\mathbf{g}(\mathbf{x}, t)$ was considered the primary object and the marginal densities followed naturally from it, according to the Fokker-Planck equation. This of course lead, for general $\mathbf{g}(\mathbf{x}, t)$, to the problem of intractable marginal densities, where marginal densities are in fact the desired output for most inference problems. In this section the priority moves to obtaining tractable (approximate) marginal densities. This can be done in one of two ways. The first approach simplifies the prior, such as linearising the dynamics in the *Extended Kalman filter* (EKF). The second approach, the one followed here, keeps the prior intact. Instead it weakens the constraints between the variational-drift and the variational-marginals by coupling them through a *projected* Fokker-Planck equation. This allows us to learn general drifts in tandem with approximate (tractable) marginals. This section combines the projection filter formulated in the previous section with the variational machinery of the general smoothing algorithm in chapter 3.

3.6.1 Surrogate-marginal free-energy approximation

For a vector of sufficient statistics $\boldsymbol{\phi} : \mathbb{R}^{d_\mathbf{x}} \to \mathbb{R}^{d_{\boldsymbol{\theta}}}$, let $S = \{p(\cdot, \boldsymbol{\gamma}), \boldsymbol{\gamma} \in \mathcal{M}\}$ denote the corresponding moment parameterised family of densities. Assume the true posterior in equation (3.11) is known, but the resulting marginal densities are intractable. Then the projection filter of section 3.4 provides a reasonable method for constructing approximate marginal-densities in the set S. If the posterior drift $\mathbf{g}(\mathbf{x}, t)$ in equation (3.11) is not known, then it needs to be learned (at least approximately). This can be done in the variational framework by minimising the free-energy in (3.5), but to compute the free-energy requires knowledge of the true marginal densities of the SDE with drift $\mathbf{g}(\mathbf{x}, t)$. As these are not available, it is reasonable to use the surrogate marginal densities of the projection filter. For a fixed drift $\mathbf{g}(\mathbf{x}, t)$, the free-energy $\mathcal{F}(\mathbf{g})$ is approximated as follows:

1. Construct surrogate marginals $\{\hat{q}(\mathbf{x}, \boldsymbol{\gamma}(t))\}_{t\in[0,T]}$ using the projection filter solving the moment evolution equation with $\boldsymbol{\gamma}(0) = \boldsymbol{\gamma}_0$,

$$\frac{\partial \boldsymbol{\gamma}(t)}{\partial t} = \left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}(\mathbf{x})] \right\rangle_{\hat{q}(\mathbf{x},\boldsymbol{\gamma}(t))}, \quad t \in [0, T]. \tag{3.76}$$

2. Approximate the energy functions $E_{sde}(t)$ and $E_{obs}(t)$ in equations (3.1) and (3.3), with

$$\hat{E}_{sde}(\boldsymbol{\gamma}(t)) = \frac{1}{2}\Big\langle ||\mathbf{g}(\mathbf{x}, t) - \mathbf{f}(\mathbf{x}, t)||^2_{\mathbf{D}(\mathbf{x},t)} \Big\rangle_{\hat{q}(\mathbf{x},\boldsymbol{\gamma}(t))} \tag{3.77}$$

$$\hat{E}_{obs}(\boldsymbol{\gamma}(t)) = \frac{1}{2}\sum_i \delta(t - t_i)\Big\langle ||\mathbf{y}_i - \mathbf{h}(\mathbf{x}, t_i)||^2_{\mathbf{R}(\mathbf{x},t_i)} \Big\rangle_{\hat{q}(\mathbf{x},\boldsymbol{\gamma}(t))}. \tag{3.78}$$

These two approximations can be inserted directly into the generalised variational smoothing algorithm of chapter 3. More precisely, define the free-energy approximation $\hat{\mathcal{F}}(\mathbf{g},\boldsymbol{\gamma})$ such that

$$\hat{\mathcal{F}}(\mathbf{g},\boldsymbol{\gamma}) = \int_0^T \hat{E}_{obs}(\boldsymbol{\gamma}(t)) + \hat{E}_{sde}(\boldsymbol{\gamma}(t))dt. \tag{3.79}$$

Assume that $\mathbf{g}(\cdot,t)$ is parameterised by some finite dimensional $\boldsymbol{\pi}(t)$. Then an approximate Lagrangian $\hat{\mathcal{L}}(\boldsymbol{\pi},\boldsymbol{\gamma},\boldsymbol{\lambda})$ can be constructed such that

$$\hat{\mathcal{L}}(\boldsymbol{\pi},\boldsymbol{\gamma},\boldsymbol{\lambda}) = \hat{\mathcal{F}}(\mathbf{g},\boldsymbol{\gamma}) - \int_0^T dt\boldsymbol{\lambda}^{\mathrm{T}}(t)\left(\frac{\partial\boldsymbol{\gamma}(t)}{\partial t} - \left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[\boldsymbol{\phi}(\mathbf{x})]\right\rangle_{\hat{q}(\mathbf{x},\boldsymbol{\gamma}(t))}\right). \tag{3.80}$$

The difference between this Lagrangian and the one in equation (3.16) is the use of the energy approximations in equations (3.77) and (3.78), instead of the exact ones in equations (3.1) and (3.3), and the use of the projected Fokker-Planck constraints in equation (3.76), instead of the full one in equation (3.15). Taking variations of (3.80) with respect to $\boldsymbol{\pi}_t$ and $\boldsymbol{\gamma}_t$ leads to essentially the same algorithm as in (3.3.1), but all expectations are now w.r.t. the projection filter $\hat{q}(\mathbf{x},\boldsymbol{\gamma}(t))$ instead of the true marginal density.

3.7 Discussion

In this chapter the variational Gaussian framework of Archambeau et al. (2007a) was extended to handle general drift functions and marginal densities parameterised by their moment vectors. The algorithm can, in principle, accommodate higher-order exponential families or mixture models. One key problem with this is the requirement of a drift function and marginal density pair, $\mathbf{g}(\mathbf{x},t)$ and $q(\mathbf{x},t)$, that can be coupled together according to equation (2.24). If this is possible, it must also be computationally feasible to compute the moments of $q(\mathbf{x},t)$ and to rebuild $q(\mathbf{x},t)$ from its moments. In the general case, matching marginal densities to drifts and vice-versa is a non-trivial task. Though recent theoretical progress has been made for exponential families (Brigo, 2000) and mixture models (Brigo et al., 2003). A lot of the difficulty in the continuous-discrete setting is not with the assimilation of data, but with capturing and describing the densities involved, whether they be true posterior smoothing densities or approximations. As the approximations become more and more complex, computational issues quickly converge in terms of difficulty to the issues faced when dealing with the true posterior. In the second half of the chapter is was shown how the strict requirement of an explicit drift-marginal pair can be relaxed. This is through the use of a projected Fokker-Planck which can be derived simply as the limit of an assumed density smoothing algorithm. These types of limiting proofs are not necessarily the most rigorous of approaches, hence the development of more principled stochastic analysis approaches to the proof of the Fokker-Planck equation (Øksendal, 2003), and the differential geometric proof of the projection filter in section 3.4.2. But they are generally easier to interpret intuitively than proofs that work completely in an infinite dimensional setting. Despite the improved flexibility that the projection method brings, it is conjectured that the family of approximate drifts and family of surrogate marginals should be chosen wisely. If the projection filter provides a poor approximation of the marginals, then the free-energy approximation may be far from the true. This in turn could lead to highly erroneous posterior-drift approximations and free-energy values.

Chapter 4

Variational drift approximations

Contribution

- This chapter makes use of the generalised framework and its projective relaxation to propose novel drift functions for use in the variational smoothing algorithm. The drift functions fall into two classes, *approximate additive controls* and *time-varying gradient systems*. The approximate additive controls play upon the duality between the optimal smoothing solution and the optimal paths generated in *optimal-control*. It is shown how VPGA can be interpreted in the setting of optimal control. The second part of the chapter introduces an alternative form of drift that focuses on regions around equilibrium points of the Fokker-Planck equation. Both new classes of drift can be combined with the projection filter of chapter 3 to keep marginal densities in a tractable class. The final part of the chapter considers the practical aspects of capturing and tracking higher order moments. A specific class of skew density is proposed and the tractability of representing and learning *skewness* in filtering and smoothing is considered.

4.1 Prior drifts with additive control

The formulation of the projection-filter weakens the restriction of requiring of an explicitly defined drift-marginal pair. This section introduces a novel posterior-drift approximation combining the general variational framework of chapter 3 with the projection-filter in chapter (3.4). It does this by augmenting the prior drift with an additive posterior control and replacing the moment constraints in chapter 3 by the weaker projected Fokker-Planck constraints in chapter (3.4). The marginal densities are therefore constrained to evolve under the projected Fokker-Planck equation and can lie in a parameterised family we choose. The additive control guides paths towards the data, and it is shown how the optimal Gaussian smoothing algorithm of Archambeau & Opper (2011) in fact does exactly this, implicitly, through its variational parameters and Lagrange multipliers. Indeed, the variational parameters have explicit forms that results in moment equations that match those of the projection filter but with additive controls in the form

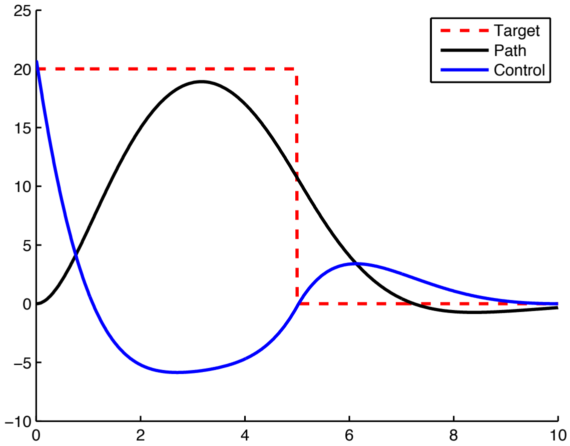


FIGURE 4.1: Illustration of optimal control problem for motor control. The optimal path is the velocity component $x_1(t)$ in black and the optimal control is the input force $u(t)$ in blue.

of additive Lagrange multipliers learned on a backward pass. This section first discuss some basic optimal control theory. It then formulates the variational smoothing problem in a similar light.

4.1.1 Stochastic optimal control

A control is a sequence of actions with future consequences. *Optimal control theory* is a mathematical formulation for using controls to achieve a specified goal. A simple deterministic example is shown in figure (4.1), in which the accelerator (blue) must be used to keep the speed of the car (black) as close as possible to the target speed (dashed red) with some cost on acceleration. The example consists of a path $\mathbf{x}(t) = (x_1(t), x_2(t)) \in \mathbb{R}^{d_\mathbf{x}}$ and a control $u(t) \in \mathbb{R}$ evolving over time with linear dynamics $\dot{x}_1 = x_2$, $\dot{x}_2 = u$ and initial condition $\mathbf{x}(0) = (0,0)^{\mathrm{T}}$, and, importantly, a cost function

$$\mathcal{C}(u) = \int_0^T (x_1 - z)^2 + u^2 dt \tag{4.1}$$

with target $z(t)$. The *optimal control* problem is to minimise the cost $\mathcal{C}(u)$ over u whilst satisfying the dynamics of the system. While the above example can be solved exactly, the optimal control problem is intractable in the general setting where dynamics can be nonlinear and stochastic. The following exposé of the general setting is adapted from Kappen (2011). Here only *finite-horizon* control problems are considered, without end-conditions and with control independent diffusions. More precisely, in the general setting the dynamics are determined by an SDE

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}, \mathbf{u}, t)dt + \sqrt{\mathbf{D}}(\mathbf{x}, t)d\mathbf{w}_t, \tag{4.2}$$

for some control $\mathbf{u}(t)$ taking values in $\mathbb{R}^{d_\mathbf{u}}$, and a cost function

$$\mathcal{C}(\mathbf{u}) = \left\langle \int_0^T dt E(\mathbf{x}, \mathbf{u}, t) \right\rangle \tag{4.3}$$

where the expectation is over all paths for some given initial condition. The *optimal cost-to-go*, $J(\mathbf{x}, t)$, satisfies the *Stochastic Hamilton-Jacobi-Bellman* (HJB) Equation

$$-\partial_t J(\mathbf{x}, t) = \min_{\mathbf{u}} \left\{ E(\mathbf{x}, \mathbf{u}, t) + \overleftarrow{\mathcal{K}}_{\mathbf{f}(\mathbf{x},\mathbf{u},t)}[J(\mathbf{x}, t)] \right\}, \tag{4.4}$$

where $\overleftarrow{\mathcal{K}}_{\mathbf{f}(\mathbf{x},\mathbf{u},t)}[\cdot]$ is given by

$$\overleftarrow{\mathcal{K}}_{\mathbf{f}(\mathbf{x},\mathbf{u},t)}[\phi(\mathbf{x})] = \mathbf{f}^{\mathrm{T}}(\mathbf{x}, \mathbf{u}, t)\nabla[\phi(\mathbf{x})] + \frac{1}{2}\mathrm{tr}\Big\{\mathbf{D}(\mathbf{x}, t)\big(\nabla\nabla^{\mathrm{T}}\big)[\phi(\mathbf{x})]\Big\}. \tag{4.5}$$

Now assume that $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ is additive-linear in $\mathbf{u}$, i.e.

$$\mathbf{f}(\mathbf{x}, \mathbf{u}, t) = \mathbf{f}(\mathbf{x}, t) + \mathbf{u}, \tag{4.6}$$

and assume the energy $E(\mathbf{x}, \mathbf{u}, t)$ is quadratic in $\mathbf{u}$, i.e.

$$E(\mathbf{x}, \mathbf{u}, t) = V(\mathbf{x}, t) + \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{x}, t)\mathbf{u}. \tag{4.7}$$

The system dynamics and cost are arbitrarily complex, but the control dynamics and cost are linear and quadratic respectively. This is the special case considered in Kappen (2005a,b) where a *path integral* formulation is possible. For this case, the HJB equation reduces to

$$-\partial_t J(\mathbf{x}, t) = \min_{\mathbf{u}} \left\{ V(\mathbf{x}, t) + \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{x}, t)\mathbf{u} + \overleftarrow{\mathcal{K}}_{\mathbf{f}(\mathbf{x},t)+\mathbf{u}}[J(\mathbf{x}, t)] \right\}. \tag{4.8}$$

This quadratic form can be minimised w.r.t. $\mathbf{u}$ to give

$$\mathbf{u}(\mathbf{x}, t) = -\mathbf{D}(\mathbf{x}, t)\nabla_{\mathbf{x}} J(\mathbf{x}, t). \tag{4.9}$$

This defines the optimal control for each $\mathbf{x}$ and t. Inserting $u(\mathbf{x}, t)$ back into the the HJB equation gives

$$-\partial_t J(\mathbf{x}, t) = V(\mathbf{x}, t) - \frac{1}{2}\nabla_{\mathbf{x}}^{\mathrm{T}} J(\mathbf{x}, t)\mathbf{D}(\mathbf{x}, t)\nabla_{\mathbf{x}} J(\mathbf{x}, t) + \overleftarrow{\mathcal{K}}_{\mathbf{f}(\mathbf{x},t)}[J(\mathbf{x}, t)]. \tag{4.10}$$

Substituting $J(\mathbf{x}, t) = -\log\psi(\mathbf{x}, t)$ yields

$$V(\mathbf{x}, t)\psi(\mathbf{x}, t) = \big(\partial_t + \overleftarrow{\mathcal{K}}_{\mathbf{f}}\big)[\psi(\mathbf{x}, t)]. \tag{4.11}$$

The particular control problem, i.e. finite horizon and no end conditions, and some of the variables used, i.e. $\mathbf{D}_t^{-1}$ for the quadratic cost, were chosen retrospectively, but this doesn't detract from the fact that equation (4.11) is equivalent to equation (3.10) for $V(\mathbf{x},t) = U(\mathbf{x},t)$, and equations (4.9) and (4.6) are, together, equivalent to equation (3.11). Connections between estimation and optimal control have received recent attention in Kappen et al. (2009), Kappen (2011), Todorov (2008). From the above exposé, I would conjecture that optimal control is a more general framework that subsumes estimation. Both can be formulated in a variational framework and cost functions are generalisations of free-energies. The appeal of this "duality" for us, is that any progress made with the general variational algorithm of chapter (3) is likely to be transferable to problems in optimal control. Note that the variational algorithms are not applicable in control problems with state constraints because the true posterior has smaller support than the approximation (Mensink et al., 2010).

4.1.2 Additive drift control

To strengthen the connection between estimation and control is illuminative to make the argument in reverse. The posterior drift given in equation (3.11) derived from exact inference can be written in an additive control form

$$\mathbf{g}(\mathbf{x}, y) = \mathbf{f}(\mathbf{x}, t) + \mathbf{u}(\mathbf{x}, t) \tag{4.12}$$

where $\mathbf{u} : \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}_+ \to \mathbb{R}^{d_c}$ defines a control vector. Let Q denote the measure over paths (with respect to the Wiener measure) generated according to the SDE in equation (2.21) with drift $\mathbf{g}(\mathbf{x}, t)$ given by equation (4.12). Let $q(\mathbf{x}, t)$ denote the marginals of Q at time t. The relative entropy KL$[Q||P_{prior}]$ in equation (3.2) is then the integral of a free-energy function $E_{sde}(t)$ in equation (3.1) given by

$$E_{sde}(t) = \frac{1}{2}\Big\langle ||\mathbf{u}(\mathbf{x}, t)||^2_{\mathbf{D}(\mathbf{x},t)} \Big\rangle_{q(\mathbf{x},t)}. \tag{4.13}$$

Thus, the prior drift $\mathbf{f}(\mathbf{x}, t)$ has been removed from the free-energy and the resulting KL term is simply a quadratic form of the control vector $\mathbf{u}(\mathbf{x}, t)$. Inserting this into the exact Lagrangian in equation (3.9) with $U(\mathbf{x}, t) = V(\mathbf{x}, t)$, and taking independent variations w.r.t. q and $\mathbf{u}$ leads, as expected, to the solution $\mathbf{u}(\mathbf{x}, t) = \mathbf{D}(\mathbf{x}, t)\nabla \ln \psi(\mathbf{x}, t)$. While estimation can be considered a special case of optimal control when inference is exact. This does not mean the same relation holds for approximate inference and approximate control. In exact inference, both schemes attempt to learn a "posterior" drift using a cost function that weights paths generated by equation (2.21) using the corresponding drift. In the approximate inference setting, we can assume that both schemes attempt to learn an approximate drift. For us, this means it is possible to approach approximate inference in a control setting by simply augmenting the prior drift $\mathbf{f}(\mathbf{x}, t)$ with a suboptimal control $\hat{\mathbf{u}}(\mathbf{x}, t)$ which may be, for example, linear in $\mathbf{x}$. This would result in a free-energy that is quadratic in $\hat{\mathbf{u}}(\mathbf{x}, t)$. It appears that this is fundamentally different from the variational framework in chapter (3), in that the variational drift $\mathbf{g}(\mathbf{x},t) = \mathbf{f}(\mathbf{x},t) + \hat{\mathbf{u}}(\mathbf{x},t)$ contains the prior. This would cause significant problems in the general framework of chapter (3) because we would expect the variational drift to produce marginal densities with intractable forms. These problems can be remedied through the use of the projected Fokker-Planck equation in section (3.6.1). The control $\hat{\mathbf{u}}(\mathbf{x},t)$ can be used to keep the marginal densities close to a tractable manifold and guide them towards the data. Though this approximation scheme appears different from the VGPA algorithm of section (3.3.2). It can be shown how the VGPA algorithm takes this exact form. Let us assume that $\mathbf{D}(\mathbf{x},t) = \mathbf{D}(t)$. In the variational GP setting of Archambeau & Opper (2011) and section (3.3.2), assuming state independent diffusions and certain positive-definiteness and drift constraints are satisfied, the variational parameters $\mathbf{A}(t)$ and $\mathbf{b}(t)$ can be written in explicit forms

$$\mathbf{A}(t) = \big\langle \nabla_{\mathbf{x}} \mathbf{f}^{\mathrm{T}}(\mathbf{x},t) \big\rangle_{q(\mathbf{x},t)} - 2\mathbf{D}(t)\boldsymbol{\Lambda}(t) \tag{4.14}$$

$$\mathbf{b}(t) = \big\langle \mathbf{f}(\mathbf{x},t) \big\rangle_{q(\mathbf{x},t)} - \mathbf{A}(t)\boldsymbol{\mu}(t) - \mathbf{D}(t)\boldsymbol{\nu}(t) \tag{4.15}$$

where $q(\mathbf{x},t)$ denote the marginal densities of the corresponding Gaussian process and $\boldsymbol{\nu}(t)$ and $\boldsymbol{\Lambda}(t)$ denote the Lagrange multipliers corresponding to the moments constraints in equations (3.22) and (3.23). Inserting equations (4.14) and (4.15) back into equations (3.22) and (3.23) yields

$$\dot{\boldsymbol{\mu}}(t) = \big\langle \mathbf{f}(\mathbf{x},t) \big\rangle_{q_t} - \mathbf{D}(t)\boldsymbol{\nu}(t) \tag{4.16}$$

$$\dot{\boldsymbol{\Sigma}}(t) = \big\langle \nabla_{\mathbf{x}} \mathbf{f}^{\mathrm{T}}(\mathbf{x},t) \big\rangle_{q_t} \boldsymbol{\Sigma}(t) + \boldsymbol{\Sigma}(t) \big\langle \nabla_{\mathbf{x}} \mathbf{f}^{\mathrm{T}}(\mathbf{x},t) \big\rangle_{q_t} + \mathbf{D}(t) - 4\mathbf{D}(t)\boldsymbol{\Lambda}(t)\boldsymbol{\Sigma}(t). \tag{4.17}$$

These are exactly the GP projection-filter equations in section 3.4.4.1, but with additional controls $\mathbf{u}_{\boldsymbol{\mu}}(t) := \mathbf{D}(t)\boldsymbol{\nu}(t)$ and $\mathbf{u}_{\boldsymbol{\Sigma}}(t) := 4\mathbf{D}(t)\boldsymbol{\Lambda}(t)\boldsymbol{\Sigma}(t)$. In the VGPA algorithm, the variational parameters account for the constraints and the observations. In this *new* light they act as controls, both guiding the mean towards the data and updating the covariance, and ensuring the approximate marginal densities remain on the Gaussian manifold. In figure 4.2 a simple toy example shows how the VGPA algorithm can be applied directly to a control problem adapted from the double-well inference problem. A target $z(t)$ flips between one of two states $\{\pm 1\}$, and an aim $x(t)$ can move and be calibrated to the accuracy of a double-well equilibrium distribution with a peak around each of the two possible states. As the target appears, the aim must move as close to the target as possible in a mean-squared sense. The position of the target $z(t)$ can only be detected at intermittent time intervals, as would be the case with a digital sensor. This problem cannot be solved exactly using optimal-control theory because of the nonlinear dynamics of the double well, but it can be solved approximately by the VGPA algorithm. In the example, the aim (black) wrongly guesses the initial state of the target. The control (blue), which in the VGPA algorithm is the negation of the Lagrange multiplier $\nu(t)$, corrects this by taking a highly negative initial value; forcing the aim across to the other well. The “squiggles”

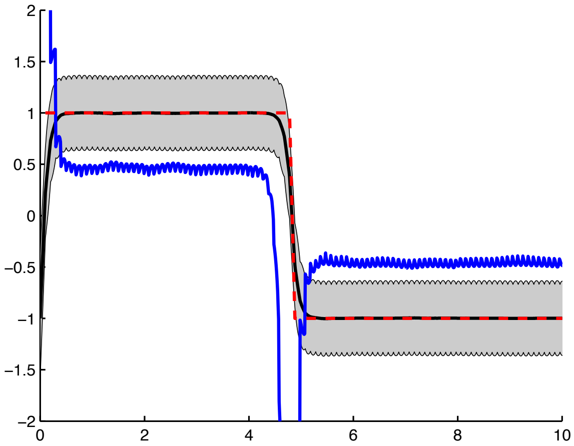


FIGURE 4.2: Plot of double-well target-practice. Target $z(t)$ (dashed red), mean path $\mu(t)$ (black), confidence region (shaded), and Lagrange multiplier $\nu(t)$ (blue).

in $\nu(t)$ in between the sensor measurements are due to the mean value of the aim trying to revert to the mean value of the double-well equilibrium distribution. The control $\nu(t)$ stops this when the target is detected, keeping the aim on the correct state until the target is about to jump to the alternative state and $\nu(t)$ compensates. It is not a "real-time" control problem, because the control $\nu(t)$ anticipates the movement of the target, but it is an optimal Gaussian solution to the control problem in a free-energy sense.

4.1.3 Motion equations

In the example in the previous section, the control $\boldsymbol{\nu}(t)$ lay in the "velocity" space of the system. This is a general feature of the Lagrange multipliers $\boldsymbol{\nu}(t)$ and $\boldsymbol{\Sigma}(t)$ in equations (4.16) and (4.17), where they occupy the same space as $\dot{\boldsymbol{\mu}}(t)$ and $\dot{\boldsymbol{\Sigma}}(t)$. In many control problems with a physical interpretation, the control lies in the "acceleration" space of the system where they can act as external forces. Assuming still the form $\mathbf{D}(\mathbf{x}, t) = \mathbf{D}(t)$, systems with Wiener fluctuations acting as forces, can be represented through SDEs of the form

$$
\begin{aligned}
d\mathbf{x}_t &= d\mathbf{v}_t dt && (4.18)\\
d\mathbf{v}_t &= \mathbf{f}(\mathbf{x}_t, \mathbf{v}_t, t)dt + \sqrt{\mathbf{D}}(t)d\mathbf{w}_t, && (4.19)
\end{aligned}
$$

where $\mathbf{x}_t$ is a position vector, $\mathbf{v}_t$ a velocity vector, and $d\mathbf{w}_t$ the standard Wiener process in $\mathbb{R}^{d_\mathbf{v}}$. Paths $\mathbf{x}_t$ are now once differentiable, in contrast to the nowhere-differentiable paths of $\mathbf{v}_t$. This formulation matches a noisy, nonlinear version of the motor control problem in figure 4.1, as well as animal locomotion models (Collins & Stewart, 1993) and other Dynamic Movement Primitives (DMPs) for human and robot motion (Schaal, 2006). The lack of system noise in the deterministic relation between $d\mathbf{x}$ and $d\mathbf{v}$ means that a free-energy approach is not directly applicable. This does not pose a problem, because we are not attempting to approximate this relation. Indeed, for an approximating process Q with drift $\mathbf{g}(\mathbf{x}_t, \mathbf{v}_t, t)$, $d\mathbf{x}_t = d\mathbf{v}_t dt$, the relative entropy between the prior process P_{prior} with drift $\mathbf{f}(\mathbf{x}_t, \mathbf{v}_t, t)$ and the approximation Q is simply

$$\mathrm{KL}[Q||P_{prior}] = \frac{1}{2}\int_0^T dt \Big\langle ||\mathbf{g}(\mathbf{x}, \mathbf{v}, t) - \mathbf{f}(\mathbf{x}, \mathbf{v}, t)||^2_{\mathbf{D}(t)} \Big\rangle_{q(\mathbf{x},\mathbf{v},t)}, \tag{4.20}$$

where $q(\mathbf{x}, \mathbf{v}, t)$ denotes the approximating marginal density at time t (appendix A.3.4). While higher-order fluctuations such as “jerk” can be incorporated in a similar fashion, the dimension will increase with every new generalised motion. A bi-product of the introduction of the higher order motions is the smoothing effect on the motion of the position variable $\mathbf{x}_t$. Indeed, the classical approach to incorporating *coloured* noise into state-space models and filters uses a similar method (Anderson & Moore, 1979). Explicit examples of this are presented in section 5.3.1. Continuing the idea of “duality”, we could conjecture a direct link between solutions to optimal-control control problems involving higher-order derivatives of motion and inference in estimation problems involving time-correlated noise.

4.2 Gradient systems

Gradients systems, with constant diffusions, provide particular drifts for which it is possible to obtain explicit forms for the steady-state distribution under the Fokker-Planck equation. They are well established in physics, control, and mathematics (Caughey & Ma, 1982, Fuller, 1969, Prato, 2006). This section introduces an alternative form of drift that focuses on regions around equilibrium points of the Fokker-Planck equation. Time-varying gradient systems are used to build well behaved vector fields on the space of probabilities. These too can be combined with the projection filter of the previous section to keep marginal densities in a tractable family. In this section it is assumed that $\mathbf{D}(\mathbf{x}, t) = \epsilon\mathbf{I}$ for some $\epsilon > 0$, though the exposition is extendable to any positive definite constant matrix $\mathbf{D}$.

4.2.1 Steady-state potentials

In this section it is assumed that the drift $\mathbf{g}(\mathbf{x}, t) = \mathbf{g}(\mathbf{x})$ is time-invariant. Steady-steady or *equilibrium* distributions form an integral part of any analysis involving the Fokker-Planck equation. They constitute the equilibrium points of the Fokker-Planck equation and (equivalently) form the ground-state eigenfunctions of the Fokker-Planck operator with zero eigenvalue. More precisely, let $p_0(\mathbf{x})$ denote the probability density function that solves the Fokker-Planck (ground-state) equation

$$\overrightarrow{\mathcal{K}}_{\mathbf{g}}[p(\mathbf{x}, t)] = 0 \tag{4.21}$$

where, for some $\epsilon > 0$ and drift $\mathbf{g}(\mathbf{x})$, $\overrightarrow{\mathcal{K}}_{\mathbf{g}}$ denotes the Fokker-Planck operator

$$\overrightarrow{\mathcal{K}}_{\mathbf{g}}[p(\mathbf{x})] = -\nabla^{\mathrm{T}}\left[p(\mathbf{x})\mathbf{g}(\mathbf{x})\right] + \epsilon\frac{1}{2}\mathrm{tr}\Big\{\left(\nabla\nabla^{\mathrm{T}}\right)\left[p(\mathbf{x})\right]\Big\}. \tag{4.22}$$

The conditions $p \geq 0$ and $\int p d\mathbf{x} = 1$ are assumed to be implicit. In most applications involving gradient systems, the drift $\mathbf{g}(\mathbf{x})$ is given and the problem centres around finding the steady-state $p_0(\mathbf{x})$. In this thesis a different situation presents itself. The variational algorithm learns drift functions and marginal densities in tandem. It can either restrict the class of drift functions (e.g. to be linear) or it can restrict the class of marginal densities (e.g. to be Gaussian). But no matter which part of the model is restricted (drift or marginal), the other part of the model must follow suit because the variational algorithm requires a drift-marginal pairing that is at least close enough for the projection filter to be efficient. With this in mind, let us, unconventionally, assume $p_0(\mathbf{x})$ is known and solve (4.21) for $\mathbf{g}(\mathbf{x})$. It is not difficult to show the solution, which will be denoted $\mathbf{g}_0(\mathbf{x})$, is given by

$$\mathbf{g}_0(\mathbf{x}) = \epsilon \frac{1}{2} \nabla \log p_0(\mathbf{x}). \tag{4.23}$$

Indeed, for completeness,

$$\begin{aligned}
\vec{\mathcal{K}}_{\mathbf{g}_0}[p_0(\mathbf{x})] &= -\nabla^{\mathrm{T}}\left[p_0(\mathbf{x})\mathbf{g}_0(\mathbf{x})\right] + \epsilon\frac{1}{2}\mathrm{tr}\Big\{\left(\nabla\nabla^{\mathrm{T}}\right)\left[p_0(\mathbf{x})\right]\Big\} && (4.24)\\
&= -\epsilon\frac{1}{2}\nabla^{\mathrm{T}}\left[p_0(\mathbf{x})\nabla\log p_0(\mathbf{x})\right] + \epsilon\frac{1}{2}\mathrm{tr}\Big\{\left(\nabla\nabla^{\mathrm{T}}\right)\left[p_0(\mathbf{x})\right]\Big\} && (4.25)\\
&= -\epsilon\frac{1}{2}\nabla^{\mathrm{T}}\left[\nabla p_0(\mathbf{x})\right] + \epsilon\frac{1}{2}\mathrm{tr}\Big\{\left(\nabla\nabla^{\mathrm{T}}\right)\left[p_0(\mathbf{x})\right]\Big\} && (4.26)\\
&= 0. && (4.27)
\end{aligned}$$

For cleanliness, let us move the $\frac{1}{2}$ into the constant ϵ. Recapping the above, a drift-marginal pairing $(\mathbf{g}_0, p_0)$ has been defined that satisfies the ground-state version of the Fokker-Planck equation. For example, a quick way to construct a time-invariant linear model is to insert a Gaussian potential into (4.23), i.e. to define a drift

$$\begin{aligned}
\mathbf{g}(\mathbf{x}) &:= \nabla \log \mathcal{N}_{\mathbf{x}}(\mathbf{m}, \mathbf{C}) && (4.28)\\
&= -\frac{1}{2}\nabla||\mathbf{x}-\mathbf{m}||^2_{\mathbf{C}} && (4.29)\\
&= -\mathbf{C}^{-1}\mathbf{x} + \mathbf{C}^{-1}\mathbf{m} && (4.30)\\
&= \mathbf{A}\mathbf{x} + \mathbf{b} && (4.31)
\end{aligned}$$

where $\mathbf{A} := -\mathbf{C}^{-1}$ and $\mathbf{b} := \mathbf{C}^{-1}\mathbf{m}$ denote *canonical* drift-parameters. By definition, $\mathbf{A}$ is negative definite and $\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x})$ defines a stable (deterministic) system.

4.2.2 Parameterised vector fields

Let $S = \{p_{\gamma}(\cdot), \gamma \in \Omega\}$ denote a parameterised family of densities for some $\Omega \subset \mathbb{R}^n$. Using the set S, let us generate a set of drift functions $\mathcal{G}(S) := \{\mathbf{g}_{\boldsymbol{\pi}}(\cdot), \boldsymbol{\pi} \in \Omega\}$ such that

$$\mathbf{g}_{\boldsymbol{\pi}}(\cdot) := \epsilon \nabla_{\mathbf{x}} \log p_{\boldsymbol{\pi}}(\cdot), \quad \forall \mathbf{g} \in \mathcal{G}(S), \epsilon > 0. \tag{4.32}$$

In words, a set of drift functions $\mathcal{G}(S)$ has been created from a set of densities S such that each $p_{\boldsymbol{\pi}} \in S$ is the equilibrium distribution of a drift $\mathbf{g}_{\boldsymbol{\pi}} \in \mathcal{G}(S)$, for some temperature parameter ϵ. Though it looks similar to the backward map between moment and dual parameters in equation (A.8), it is important not to confuse the two. The derivative in (4.32) is with respect to $\mathbf{x}$, and both $p_{\boldsymbol{\pi}}(\cdot)$ and $\mathbf{g}_{\boldsymbol{\pi}}(\cdot)$ are maps on $\mathbf{x}$. The Fokker-Planck equation defines a vector field on the space of probability density functions. Regions around the equilibrium points of the vector field are of importance, in a similar way to finite dimensional differential analyses. Changing $\boldsymbol{\pi}$ affects the flow of the vector field implied by the Fokker-Planck equation. But as long as $\boldsymbol{\pi}$ remains in Ω, then $\mathbf{g}_{\boldsymbol{\pi}} \in \mathcal{G}(S)$ will always have an equilibrium distribution $p_{\boldsymbol{\pi}} \in S$. The above analysis is not sufficient to solve the drift-marginal pairing problem for the variational smoother. This is because the presence of data will always result in a non stationary posterior distribution, and thus any credible approximation should consist of a time-varying drift function to allow for representation of the data.

4.2.3 Temporal-variant drifts

Assume an initial density $p_0(\mathbf{x})$ is given and let us project $p_0(\mathbf{x})$ onto S to give an initial density $p_{\boldsymbol{\gamma}(0)}(\mathbf{x})$ with parameter $\boldsymbol{\gamma}(0) \in \Omega$. From the previous section, we know the pair $(\mathbf{g}_{\boldsymbol{\gamma}(0)}, p_{\boldsymbol{\gamma}(0)})$ satisfy the functional

$$\overrightarrow{\mathcal{K}}_{\mathbf{g}_{\boldsymbol{\gamma}(0)}}[p_{\boldsymbol{\gamma}(0)}] = 0. \tag{4.33}$$

Now assume that we are presented with a time-varying parameter $\boldsymbol{\pi}(t)$ such that $\boldsymbol{\pi}(t) \in \Omega$ for all $t \in [0, T]$ and $\boldsymbol{\pi}(0) = \boldsymbol{\gamma}(0)$. Starting with initial condition $p(\mathbf{x}, 0) = p_{\boldsymbol{\gamma}(0)}(\mathbf{x})$, consider solutions to the Fokker-Planck equation

$$\left(\partial_t - \overrightarrow{\mathcal{K}}_{\mathbf{g}_{\boldsymbol{\pi}(t)}}\right)[p(\mathbf{x}, t)] = 0. \tag{4.34}$$

Obviously the initial condition $p(\mathbf{x}, 0) = p_{\boldsymbol{\gamma}(0)}(\mathbf{x})$ lies in S, and for any future time $t > 0$ it holds that $\mathbf{g}_{\boldsymbol{\pi}(t)}$ has an equilibrium density in S. Therefore it is reasonable to assume that the solution $p(\mathbf{x}, t)$ remains in S, or at least close to S, for all $t \in [0, T]$. To make sure of this the projected Fokker-Planck equation in section 3.4.2 can be used. Any solutions that drift off S are instantly projected back on. If it helps, the reader can consider a finite dimensional analogy. Consider the differential equation

$$\frac{d\mathbf{x}}{dt} = \mathbf{x}(t) - \boldsymbol{\pi}(t) \tag{4.35}$$

where $\mathbf{x}(t)$ is a position vector and $\boldsymbol{\pi}(t) \in V$ is drift parameter lying in some subspace V for all $t \in [0, T]$. With the initial condition $\mathbf{x}(0) = \boldsymbol{\pi}(0)$, the systems begins in equilibrium with $\mathbf{x}(0) \in V$. As time evolves, $\boldsymbol{\pi}(t)$ moves, but if $\boldsymbol{\pi}(t)$ remains in V then so does $\mathbf{x}(t)$. This is because the velocity vector $\mathbf{x}(t) - \boldsymbol{\pi}(t)$ lies in V for all $t \in [0, T]$. For the more general case we do not have the luxury of a Euclidean subspace V, but instead a submanifold. The use of the projection filter helps ensure that the solutions remain in the desired space.

This shows similarities to the ensemble method used in Friston (2008b). The difference is that here the a drift is designed to generate a fixed form of marginal density and in Friston (2008b) an arbitrary form of marginal density is used to guide an ensemble.

4.3 Tracking higher order moments

Let us return to the moment equation of (3.13). Using this equation it is possible to track the evolution of any moment, or in fact the expectation of any twice differentiable function, under the marginal density $q(\mathbf{x}, t)$, generated according to (2.21) with arbitrary drift $\mathbf{g}(\mathbf{x}, t)$. Note that the same result can be obtained using Itô's lemma (Jazwinski, 1970, lemma 4.2). Computing evolution equations for higher order moments can be quite intensive, so let us focus attention on a necessary condition to obtain a positive result with regards to the main hypothesis of the section - the ability to track properties of the third order moment or skewness. For any random vector $\mathbf{x}$ let $\mathbf{m}_1$, $\bar{\mathbf{m}}_2$, and $\bar{\mathbf{m}}_3$ denote the first moment, the second central moment, and the third central moment, respectively, given by

$$\begin{aligned} \mathbf{m}_1 &:= \langle \mathbf{x} \rangle && (4.36) \\ \bar{\mathbf{m}}_2 &:= \big\langle (\mathbf{x} - \mathbf{m}_1)(\mathbf{x} - \mathbf{m}_1)^{\mathrm{T}} \big\rangle && (4.37) \\ \bar{m}_3^{(ijk)} &:= \big\langle (x^{(i)} - m_1^{(i)})(x^{(j)} - m_1^{(j)})(x^{(k)} - m_1^{(k)}) \big\rangle, \quad i, j, k \in \mathbb{N}_{d_{\mathbf{x}}}. && (4.38) \end{aligned}$$

Element notation is used for the third central moment to avoid the need for tensor notation. Now assume that the random vector $\mathbf{x}$ represents the position $\mathbf{x}(t)$ at time t of paths generated according to (2.21). The following proposition is well known, but generally only computed up to the first and second moment (Jazwinski, 1970, Maybeck, 1979).

Proposition 4.1. *Let $p(\mathbf{x}, t)$ be the one-time marginal density generated according to SDE* (2.21) *with drift $\mathbf{g}(\mathbf{x}, t)$ and diffusion map $\mathbf{D}(\mathbf{x}, t)$. Then the moments $(\mathbf{m}_1(t), \bar{\mathbf{m}}_2(t), \bar{\mathbf{m}}_3(t))$ of $p(\mathbf{x}, t)$ evolve according to*

$$\begin{aligned} \dot{\mathbf{m}}_1(t) &= \big\langle \mathbf{g}(\mathbf{x}, t) \big\rangle_{p(\mathbf{x},t)} && (4.39) \\ \dot{\bar{\mathbf{m}}}_2(t) &= \Big\langle \mathbf{g}(\mathbf{x}, t)\big(\mathbf{x} - \mathbf{m}_1(t)\big)^{\mathrm{T}} + \big(\mathbf{x} - \mathbf{m}_1(t)\big)\mathbf{g}^{\mathrm{T}}(\mathbf{x}, t) + \mathbf{D}^2(\mathbf{x}, t) \Big\rangle_{p(\mathbf{x},t)} && (4.40) \\ \dot{\bar{m}}_3^{(ijk)}(t) &= \sum_{l \in \{i,j,k\}} \Bigg\langle g^{(l)}(\mathbf{x}, t) \prod_{n \in \{i,j,k\}/l} \big(x^{(n)} - m_1^{(n)}(t)\big) \Bigg\rangle \\ &\quad + \sum_{l \in \{i,j,k\}} \sum_{\substack{n \in \{i,j,k\}/l \\ w \in \{i,j,k\}/\{l,n\}}} \Big\langle D^{(nl)}(\mathbf{x}, t)\big(x^{(w)} - m_1^{(w)}(t)\big) \Big\rangle, && (4.41) \end{aligned}$$

for $i, j, k \in \mathbb{N}_{d_{\mathbf{x}}}$*, with initial conditions*

$$\begin{aligned}
\mathbf{m}_1(0) &= \langle \mathbf{x} \rangle_{p(\mathbf{x},0)} && (4.42)\\
\bar{\mathbf{m}}_2(0) &= \Big\langle \big(\mathbf{x} - \mathbf{m}_1(0)\big)\big(\mathbf{x} - \mathbf{m}_1(0)\big)^{\mathrm{T}} \Big\rangle_{p(\mathbf{x},0)} && (4.43)\\
\bar{m}_3^{(ijk)}(0) &= \Bigg\langle \prod_{l \in \{i,j,k\}} \big(x^{(l)} - m_1^{(l)}(0)\big) \Bigg\rangle_{p(\mathbf{x},0)}, \quad i, j, k \in \mathbb{N}_{d_{\mathbf{x}}}. && (4.44)
\end{aligned}$$

The proof of this is given in appendix A.3.5. It is important to note that $p(\mathbf{x}, t)$ is the true marginal density generated by $\mathbf{g}(\mathbf{x}, t)$. So for general drifts, $p(\mathbf{x}, t)$ will be characterised by higher order moments than the first three presented here, and this is merely a snap-shot of the low order statistics of $p(\mathbf{x}, t)$. Most importantly, to compute the expectations in the proposition requires knowing the *full* marginal density $p(\mathbf{x}, t)$. Luckily for us, the projection filter approximation in section (3.5.2) allows us to simply replace $p(\mathbf{x}, t)$ with an approximation $q(\mathbf{x}, t)$. This is also the heuristic, but effective, approach of assumed density filtering (Kushner, 1967). If we assume $q(\mathbf{x}, t)$ to be a normal density, which will be denoted $q_N(\mathbf{x}|\boldsymbol{\mu}(t), \boldsymbol{\Sigma}(t))$, then we can utilise the relations $\boldsymbol{\mu} = \mathbf{m}_1$ and $\boldsymbol{\Sigma} = \bar{\mathbf{m}}_2$, and obtain assumed Gaussian filtering solutions in terms of evolution equations for $\boldsymbol{\mu}(t)$ and $\boldsymbol{\Sigma}(t)$, identical to the projection filter evolution equations in section 3.4.4.1. Projection onto the Gaussian manifold constrains $\bar{\mathbf{m}}_3$ to be a zero tensor for all time and likewise $\dot{\bar{\mathbf{m}}}_3$. Therefore $\dot{\bar{\mathbf{m}}}_3$ does not feature in the GP projection filter evolution equations. The rest of the section considers how the evolution of the third order cumulant can be used and incorporated into filtering and smoothing.

4.3.1 Multivariate skew-normal distribution

The skew normal distribution is an extension to the normal distribution that allows odd-ordered moments and cumulants to be captured through a *shape* or *skewness* parameter. Its original multivariate formulation, the one used here, was introduced in Azzalini & Valle (1996). It consists of a normal (or Gaussian) distribution "reshaped" by the use of a *cumulative*-normal distribution. More precisely, let us define the density

$$p_{SN}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\gamma}) = 2\phi_{\boldsymbol{\Lambda}}(\mathbf{x} - \boldsymbol{\mu})\Phi(\boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu})) \qquad (4.45)$$

where

$$\phi_{\boldsymbol{\Lambda}}(\mathbf{x}) = |2\pi\boldsymbol{\Lambda}|^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}||\mathbf{x}||^2_{\boldsymbol{\Lambda}}\Big) \qquad (4.46)$$

and

$$\Phi(x) = (2\pi)^{-\frac{1}{2}} \int_{\infty}^{x} \exp\Big(-\frac{y^2}{2}\Big) dy. \qquad (4.47)$$

The parameters $(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\gamma})$ are referred to as the location, dispersion, and shape or skewness parameter, respectively. Note the dimensions $\boldsymbol{\Lambda} \in \mathbb{R}^{d_{\mathbf{x}} \times d_{\mathbf{x}}}$, $\boldsymbol{\mu} \in \mathbb{R}^{d_{\mathbf{x}}}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{d_{\mathbf{x}}}$. Setting $\boldsymbol{\gamma} = 0$, it holds that $\Phi(0) = \frac{1}{2}$, and the standard normal density $p_N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda})$ is recovered. One of the appealing properties of the Gaussian density is the ease at which one can move between parameters and moments. Indeed, in the Gaussian setting they are one in the same, with the identities $\boldsymbol{\mu} = \mathbf{m}_1$ and $\boldsymbol{\Sigma} = \bar{\mathbf{m}}_2$ being used in the GP projections. For the skew-normal density, things get slightly more involved. To use skew-normal densities in a filtering algorithm, it must be possible to reconstruct the parameters of a skew-normal density from its set of finite-order moments, in an analogous fashion to the way the Gaussian is characterised by $\mathbf{m}_1$ and $\bar{\mathbf{m}}_2$. This can be done through the use of the cumulant generating function (CGF). The CGF, $\mathcal{K}_{SN}(\mathbf{z})$, for the multivariate skew-normal density $p_{SN}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\gamma})$ is given by

$$\mathcal{K}_{SN}(\mathbf{z}) = \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z} + \frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{z} + \log\left(2\Phi(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z})\right) \tag{4.48}$$

where the vector $\boldsymbol{\delta}$ is given by

$$\boldsymbol{\delta} = \frac{\boldsymbol{\Lambda}\boldsymbol{\gamma}}{\sqrt{1+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\gamma}}}. \tag{4.49}$$

This leads to a set of cumulants (shown in appendix A.3.8),

$$\begin{aligned}
\boldsymbol{\kappa}_1 &= \boldsymbol{\mu} + (2/\pi)^{1/2}\boldsymbol{\delta} && (4.50)\\
\boldsymbol{\kappa}_2 &= \boldsymbol{\Lambda} - (2/\pi)\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}} && (4.51)\\
\kappa_3^{(ijk)} &= (2/\pi^3)^{1/2}(4-\pi)\delta_i\delta_j\delta_k, \quad i,j,k \in \mathbb{N}_{d_{\mathbf{x}}}. && (4.52)
\end{aligned}$$

Importantly, the first three cumulants of *any* distribution equate to the first moment, the second central moment, and the third central moment, respectively, (Abramowitz & Stegun, 1964, 26.1.13). Thus, in an analogous fashion to the Gaussian relations $\boldsymbol{\mu} = \mathbf{m}_1$ and $\boldsymbol{\Sigma} = \bar{\mathbf{m}}_2$, we can compute the parameters $(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\gamma})$ of a skew-normal density $p_{SN}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\gamma})$ from the moments $(\mathbf{m}_1, \bar{\mathbf{m}}_2, \bar{\mathbf{m}}_3^{diag})$ using the identities

$$\begin{aligned}
\boldsymbol{\delta} &= \pi^{1/2}\left(\frac{\bar{\mathbf{m}}_3^{diag}}{2^{1/2}(4-\pi)}\right)^{1/3} && (4.53)\\
\boldsymbol{\mu} &= \mathbf{m}_1 - (2/\pi)^{1/2}\boldsymbol{\delta}, && (4.54)\\
\boldsymbol{\Lambda} &= \bar{\mathbf{m}}_2 + (2/\pi)\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}} && (4.55)\\
\boldsymbol{\gamma} &= \frac{\boldsymbol{\Lambda}^{-1}\boldsymbol{\delta}}{(1-\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\delta})^{1/2}}, && (4.56)
\end{aligned}$$

where $\bar{\mathbf{m}}_3^{diag}$ denotes the diagonal of $\bar{\mathbf{m}}_3$ and the cube-root in (4.53) is taken component-wise. Note how equations (4.54)-(4.55) reduce to the Gaussian relations for $\boldsymbol{\delta} = 0$.

4.3.1.1 Why is skewness useful?

From an application perspective, the skew-normal distribution allows us to represent properties of models that the Gaussian cannot. In nonlinear filtering and smoothing, beliefs can quickly become non-Gaussian, as in the opening example in section 1.3, and additional descriptive power can enable complex posteriors to be captured more accurately. From a theoretical perspective, the skew-normal distribution will always be more optimal than the Gaussian distribution, no matter what criteria of optimality is used. This is seen simply through the relation

$$\inf_{\boldsymbol{\mu},\boldsymbol{\Lambda},\boldsymbol{\gamma}} \Omega(p_{SN}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\boldsymbol{\gamma})) \leq \inf_{\boldsymbol{\mu},\boldsymbol{\Lambda}} \Omega(p_{SN}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\mathbf{0})) = \inf_{\boldsymbol{\mu},\boldsymbol{\Lambda}} \Omega(p_{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda})) \tag{4.57}$$

for any objective Ω. With regards to more complicated distributions the skew-normal distribution retains some simplistic properties. The dimension of its parameter space is only larger than that of the Gaussian by a linear order of magnitude in $d_{\mathbf{x}}$. This is very important for high dimensional problems, where a general $d_{\mathbf{x}} \times d_{\mathbf{x}} \times d_{\mathbf{x}}$-dimensional skew-tensor makes mapping forward and backward between parameters and cumulants computationally expensive. On a similar note, it is *identifiable*, allowing us to move between unique parameters and unique distributions characterised by the first three cumulants. Skew-normal distributions are also closed under linear transformations, which proves they are also closed under marginalisation. These properties are very important for reshaping and propagating beliefs. Its cumulant generating function and its entropy have tractable forms, though not quite as user friendly as the Gaussian. Most importantly, through the use of the above properties, the skew-normal distribution can be integrated into a variational framework in a nice and complete fashion (Ormerod, 2011). Despite the above, not all the properties of the skew normal density are useful for filtering and smoothing. Most importantly, they are not closed under conditioning. This prevents direct application of the "project-condition" step in the discrete-time assumed density filter in section 2.5.2. Returning to the general setting, certain approximations have been developed to remedy the conditioning problem with generally small numerical discrepancies between the exact density and the approximation (Azzalini & Capitanio, 1999). Also, conditioning is not used explicitly in variational smoothing and it should be possible to emulate the updates in skew parameters through a set of jump equations. While there may be more suitable skewed distributions, general exponential family skewed distributions would involve difficult normalisation constants, and it can be shown (L. M. Castro & Arellano-Valle, 2008) how most skew-normal extensions are equivalent to the form given in (4.45).

4.3.2 Skewed filtering and smoothing

Assumed density filtering with Gaussian assumption has a long history dating back to Jazwinski (1970), Maybeck (1979). Assumed density filtering with a skewed assumption does not have much history at all. Julier (1998) discussed the issue of capturing skewness in filtering, with the hope of applying the *unscented transform* (UT) to the problem. Improved UTs came out of this, ones that better represent higher-order statistics, but no "higher-order filters" per say. In Naveau et al. (2005) a *skewed Kalman filter*, utilising the properties of the skew-normal distribution, was introduced. The dynamics involved are linear and restricted to ensure the filter retains a skew-normal form. In short, there does not appear to be a general skew filter in the literature that can deal with nonlinear dynamics and observation models. It is believed that this is down to a combination of several reasons. The first is that to represent skewness in a general setting requires the tracking and storing of a very large three dimensional skew-tensor. This leads to algorithms that do not scale well with dimension, as was noted in Julier (1998). The other reason is that when moving beyond the Gaussian assumption the standard approach is to use Monte Carlo methods such as the particle filters discussed in section 2.5.1. Particle filters can estimate any order of cumulant from the particle set, though the accuracy of the estimate can seriously degrade with the order of the cumulant. A third reason maybe that the Gaussian assumption is sufficient for many applications. An approximation of the mean value, and a confidence region provided by the covariance, will be enough to deal with most prediction problems. But this doesn't lessen the significance of the fact that the true filtering solution is the marginal density conditioned on the data up to that time, and that this conditional density will not be Gaussian in all nonlinear settings. While the Gaussian assumption may be sufficient for some applications, a more involved assumption should produce filter approximations that can capture the true filtering density more accurately in a nonlinear setting. Investigations into *skewed smoothing* is even less present in the literature. The above discussion highlights several important points:

1. To produce efficient skew filters and smoothers, one needs to deal with the high-dimensional skew-tensor required to represent third order cumulants.

2. For skew filters and smoothers to be used in practice, they must be at least as efficient as "state-of-the-art" particle filters and smoothers and other Monte Carlo methods.

3. Skew filters and smoothers must come with a range of numerical approximation schemes for computing expectations, as in the approximate Gaussian case.

4. Skew-normal filters and smoothers must be able to over come the lack of closure under conditioning of the family skew normal densities.

The first point can be remedied, in part, by using the skew-normal formulation in Azzalini & Valle (1996). The shape parameter γ can be computed from just the diagonal vector $\bar{\mathbf{m}}_3^{diag}$ of the third-order cumulant. This means we do not need to compute and store the whole third-order cumulant, and only require its diagonal vector to rebuild a skew-normal density, given the first and second-order cumulants as well. The second point has been tested somewhat in a variational context. In Ormerod (2011) a skew-normal variational approximation is tested against MCMC methods on *generalised linear mixed models* (Zhao et al., 2006) and linear regression with *inhomogeneous noise*. The skew-normal method is shown to be significantly faster than the MCMC method, and no reason can be seen why the same result would not occur in the filtering and smoothing setting. Indeed, approximate recursive filtering and smoothing algorithms rely upon successive approximations, either through particle approximations or tractable projections. Therefore, it is conjectured that any differences witnessed in a static setting would be magnified when the approximation methods are transferred to temporal problems with repeated approximations. For the third point, it appears that there is a relatively simple trick to make all the numerical methods used in assumed Gaussian filtering applicable to the skew-normal setting without too much additional computational cost. While the fourth point could cause significant problems at some point, it should be possible to work around it in a variational context.

4.3.2.1 Skew-GP projection filter

Starting from equations (4.53)-(4.56) we immediately encounter a problem with our current choice of parameterisation. Namely, if $\boldsymbol{\delta}(t)$ is considered a function of time we encounter significant problems for $\boldsymbol{\delta}(t) \approx 0$. More precisely, from equation (4.53) we have

$$\dot{\delta}^{(i)} = \frac{c_1 \dot{\bar{m}}_3^{(iii)}}{3\delta^2}, \quad i \in \mathbb{N}_{d_{\mathbf{x}}}, \tag{4.58}$$

using the constant $c_1 = \pi^{1/2}(2^{1/2}(4-\pi))^{-1/3}$. Equation (4.41) can be used to replace $\dot{\bar{m}}_3^{(iii)}$ with evolution equations in terms of the drift $\mathbf{g}(\mathbf{x}, t)$ and diffusion $\mathbf{D}(\mathbf{x}, t)$, but the resulting equations would be highly unstable for $\boldsymbol{\delta}$ close to zero. This problem can be remedied by keeping the evolution equations in terms of the moments $(\mathbf{m}_1, \bar{\mathbf{m}}_2, \bar{\mathbf{m}}_3^{diag})$ and re-parameterising the choice of marginal density to be $p_{SN}(\mathbf{x}|\mathbf{m}_1, \bar{\mathbf{m}}_2, \bar{\mathbf{m}}_3^{diag})$. As opposed to trying to rewrite the evolution equations (4.39)-(4.41) in terms of the *minimal sufficient* parameters $(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\delta})$. The explicit form for $p_{SN}(\mathbf{x}|\mathbf{m}_1, \bar{\mathbf{m}}_2, \bar{\mathbf{m}}_3^{diag})$ can be obtained simply by inserting equations (4.53)-(4.56) into definition (4.45). Replacing the expectations in equations (4.39)-(4.41) with expectations w.r.t. a skew-normal approximation $q_{SN}(\mathbf{x}|\mathbf{m}_1(t), \bar{\mathbf{m}}_2(t), \bar{\mathbf{m}}_3^{diag}(t))$, yields

$$\dot{\mathbf{m}}_1(t) = \big\langle \mathbf{g}(\mathbf{x}, t) \big\rangle_{q_{SN}(\mathbf{x}|\mathbf{m}_1(t), \bar{\mathbf{m}}_2(t), \bar{\mathbf{m}}_3^{diag}(t))} \tag{4.59}$$

$$\dot{\bar{\mathbf{m}}}_2(t) = \Big\langle \mathbf{M}_{\mathbf{g}}(\mathbf{x}, t) \Big\rangle_{q_{SN}(\mathbf{x}|\mathbf{m}_1(t), \bar{\mathbf{m}}_2(t), \bar{\mathbf{m}}_3^{diag}(t))} \tag{4.60}$$

$$\dot{\bar{\mathbf{m}}}_3^{diag}(t) = \Big\langle \mathbf{W}_{\mathbf{g}}^{diag}(\mathbf{x}, t) \Big\rangle_{q_{SN}(\mathbf{x}|\mathbf{m}_1(t), \bar{\mathbf{m}}_2(t), \bar{\mathbf{m}}_3^{diag}(t))} \tag{4.61}$$

where, for shorthand,

$$\mathbf{M}_{\mathbf{g}}(\mathbf{x}, t) = \mathbf{g}(\mathbf{x}, t)\big(\mathbf{x} - \mathbf{m}_1(t)\big)^{\mathrm{T}} + \big(\mathbf{x} - \mathbf{m}_1(t)\big)\mathbf{g}^{\mathrm{T}}(\mathbf{x}, t) + \mathbf{D}(\mathbf{x}, t) \tag{4.62}$$

$$W_{\mathbf{g}}^{(iii)}(\mathbf{x}, t) = 3g^{(i)}(\mathbf{x}, t)\big(x^{(i)} - m_1^{(i)}(t)\big)^2 + 3D^{(ii)}(\mathbf{x}, t)\big(x^{(i)} - m_1^{(i)}(t)\big). \tag{4.63}$$

In general, the expectations in equations (4.59)-(4.61) will have to be computed numerically. This can be done simply by moving the skewness of the marginal density from the marginal into the transformation. From the form of the skew-normal density, given in equation (4.45) as the product of a normal density and a cumulative normal distribution, this results in a Gaussian expectation of a skewed transformation, as opposed to a skew normal expectation of the original transformation. More precisely, for any function $V(\mathbf{x})$ it holds that

$$\big\langle V(\mathbf{x})\big\rangle_{p_{\mathcal{SN}}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Lambda},\mathbf{d})} = \big\langle \tilde{V}(\mathbf{x};\boldsymbol{\mu},\mathbf{d})\big\rangle_{p_{\mathcal{N}}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Lambda})} \tag{4.64}$$

where

$$\tilde{V}(\mathbf{x};\boldsymbol{\mu},\mathbf{d}) := 2V(\mathbf{x})\Phi\big(\mathbf{d}^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})\big). \tag{4.65}$$

Therefore, taking a skew-normal expectation of $V(\mathbf{x})$ is equivalent to taking a normal expectation of $\tilde{V}(\mathbf{x};\boldsymbol{\mu},\mathbf{d})$. Then any of the standard methods for approximating Gaussian expectations (discussed in section 5.4.1) can be used. While the skew projection-filter equations in (4.59)-(4.61) were designed for use in the variational smoothing algorithm of section 3.6, they can also be used as a "original" assumed density filter and be used to assimilate data. The data can be assimilated in approximately in a moment matching fashion inspired by EP. At an observation time t_i with observation model $p(\mathbf{y}_i|\mathbf{x})$, define the (possibly intractable) density

$$\tilde{p}_i(\mathbf{x}) \propto p(\mathbf{y}_i|\mathbf{x})q_{SN}\big(\mathbf{x}|\mathbf{m}_1(t_i^-),\bar{\mathbf{m}}_2(t_i^-),\bar{\mathbf{m}}_3^{diag}(t_i^-)\big). \tag{4.66}$$

Then the update equations for $\big(\mathbf{m}_1(t_i),\bar{\mathbf{m}}_2(t_i),\bar{\mathbf{m}}_3^{diag}(t_i)\big)$ are given by

$$\begin{aligned}
\mathbf{m}_1(t_i) &= \big\langle\mathbf{x}\big\rangle_{\tilde{p}_i(\mathbf{x})} && (4.67)\\
\bar{\mathbf{m}}_2(t_i) &= \Big\langle\big(\mathbf{x}-\mathbf{m}_1(t_i)\big)\big(\mathbf{x}-\mathbf{m}_1(t_i)\big)^{\mathrm{T}}\Big\rangle_{\tilde{p}_i(\mathbf{x})} && (4.68)\\
\bar{m}_3^{(iii)}(t_i) &= \Big\langle(x^{(i)}-m_1^{(i)})^3\Big\rangle_{\tilde{p}_i(\mathbf{x})}, \quad i\in\mathbb{N}_{d_{\mathbf{x}}}. && (4.69)
\end{aligned}$$

These equations could also be approximated using the previously proposed numerical method, once for the normalisation constant in (4.66) and again for each of the moments in equations (4.67)-(4.69).

4.3.3 Variational skew-GP smoothing

To use skew normal approximations for marginal densities in the variational algorithm of chapter 3, requires us to find a drift that generates skew-normal marginals. Using the gradient-system theory of section 4.2 we can define a drift with skew-normal steady-states. With a skew-normal initial state and time-varying drifts that evolve in the set of skew-normal gradient systems, the marginal densities produces under the Kolmogorov forward equation (2.24) should remain the class of skew-normal densities. If they do not, then the projection filter of section (3.6) can be used to project them down onto the skew-normal manifold. For simplicity, let us assume the diffusion matrix $\mathbf{D}(\mathbf{x},t)$ is the identity matrix times a positive constant ϵ.

4.3.3.1 Skew-normal gradient system

Let $\mathbf{g}(\mathbf{x})$ denote the drift defined by

$$\begin{aligned} \mathbf{g}(\mathbf{x}) &= \nabla_{\mathbf{x}} \log p_{SN}(\mathbf{x}|\mathbf{b}, -\mathbf{A}^{-1}, \mathbf{d}) && (4.70)\\ &= \nabla_{\mathbf{x}} \log \phi_{-\mathbf{A}^{-1}}(\mathbf{x}-\mathbf{b}) + \nabla_{\mathbf{x}} \log \Phi(\mathbf{d}^{\mathrm{T}}(\mathbf{x}-\mathbf{b})) && (4.71)\\ &= \mathbf{A}(\mathbf{x}-\mathbf{b}) + \frac{\phi(\mathbf{d}^{\mathrm{T}}(\mathbf{x}-\mathbf{b}))}{\Phi(\mathbf{d}^{\mathrm{T}}(\mathbf{x}-\mathbf{b}))}\mathbf{d}. && (4.72) \end{aligned}$$

From section 4.2, we know the drift $\mathbf{g}(\mathbf{x})$ has a steady-state in the skew-normal family. This form of drift has been studied in a more general setting with a view to finding steady-state densities for the a class of *nonlinear feedback* systems and their use to stochastic optimal control (Liberzon & Brockett, 2000). It is clear that $\mathbf{g}(\mathbf{x})$ reduces to a standard linear drift in equation (3.21) that generates Gaussian marginal densities for $\mathbf{d} = \mathbf{0}$. Let us replace $\mathbf{g}(\mathbf{x}, t)$ with a time-dependent alternative, where the time dependency comes through the parameters $(\mathbf{A}(t), \mathbf{b}(t), \mathbf{d}(t))$. We are proposing to use $\mathbf{g}(\mathbf{x}|\mathbf{A}(t), \mathbf{b}(t), \mathbf{d}(t))$ as a variational drift approximation, and q_{SN} as a skew-normal marginal density approximation, in the variational smoothing framework of chapter 3. This requires q_{SN} to satisfy the moment equations (4.67)-(4.69) for $\mathbf{g}(\mathbf{x}, t) = \mathbf{g}(\mathbf{x}|\mathbf{A}(t), \mathbf{b}(t), \mathbf{d}(t))$. Inserting $\mathbf{g}(\mathbf{x}, t)$ into equation (4.59) yields

$$\begin{aligned} \dot{\mathbf{m}}_1(t) &= \big\langle \mathbf{g}(\mathbf{x}, t) \big\rangle_{q_{SN}(\mathbf{x},t)} && (4.73)\\ &= \mathbf{A}(t)\big(\mathbf{m}_1(t) - \mathbf{b}(t)\big) + \mathbf{d}(t)\Big\langle \tilde{\phi}\Big(\mathbf{d}^{\mathrm{T}}(t)\big(\mathbf{x}-\mathbf{b}(t)\big)\Big)\Big\rangle_{q_{SN}(\mathbf{x},t)} && (4.74) \end{aligned}$$

where

$$\tilde{\phi}(x) := \frac{\phi(x)}{\Phi(x)}. \tag{4.75}$$

We see for $\mathbf{d} = \mathbf{0}$ how equation (4.74) reduces to the Gaussian mean evolution equation (3.22) in the original variational Gaussian algorithm of section 3.3.2. The expectation in (4.74) is just a one-dimensional integration problem and can therefore be approximated relatively easily using the same ideas as in section 4.3.2.1. Though more involved, the other evolution equations can be derived in a similar way.

4.4 Discussion

The connection between exact inference and optimal control is simple to show in a variational formulation. The connection between variational approximate inference and suboptimal control is not so obvious. In the first part of this chapter, section 4.1 showed how the mean and covariance evolution equations of the VGPA algorithm of Archambeau et al. (2007a) can be derived in the language of optimal control. In summary:

- Using a linear drift approximation and exact Gaussian marginal constraints in the variational smoothing algorithm of section (3.3.2), is equivalent to equipping the prior drift with an approximate additive control and projecting the marginal constraints onto the Gaussian manifold.

In exact inference it is the Lagrange multiplier $\rho(\mathbf{x}, t)$ in equation (3.9) for the Fokker-Planck equation constraint, corresponding to the optimal cost-to-go $J(\mathbf{x}, t)$, that becomes the additive control. What we see from equations (4.16) and (4.17), is that in the variational GP approximation it is the Lagrange multipliers for the weakened moment constraints in equation (3.22) and (3.23), the projected Fokker Planck constraints, that become the controls in the suboptimal variational GP approximation. This connection has not been made before. It should extend to more general families and future work would look at extending the variational smoothing algorithm to general control problems.

Gradient systems are common in many areas of physics, mathematics and engineering. This is mainly due to the simple expression of there steady-state distributions. The usual problem we are confronted with when dealing with diffusion processes is that we are presented with a drift function and diffusion matrix and the goal is find either the marginal densities at particular times of the equilibrium density of the system. In this chapter, in section 4.2 the situation is reversed. In the variational smoothing algorithm it is simple to choose an approximating class for the marginal densities of the smoothing posterior, but it is difficult to find a form of drift that generates marginals in this class. Through the use of gradient systems, a range of well behaved drift functions can be derived for which we know the steady states. Each drift can be used to parameterise the Kolmogorov forward equation. The approximating drift must evolve over time to be able to be able to approximate the nonstationary nature of the posterior that follows from assimilating data. As the variational drift changes, so does the vector field over the space of probability densities and the marginal densities flow towards the equilibrium density of the corresponding drift. The unknown drift and marginal densities can be coupled together using the projected Fokker-Planck equation in the variational smoother and, in theory, they can be learned in tandem.

The skew-normal density represents a simple deviation from normality, whilst retaining many desirable properties. In section 4.3 it was shown how properties of the third order cumulant can be tracked in the setting of diffusion processes. Using the ideas of the projection filter in section 3.5.2, the evolution of an intractable posterior can be projected down onto its first three moments. It is shown how the skew-normal density can be parameterised fully by its first three moments, and this leads to it being a natural choice for non-Gaussian filtering and smoothing in continuous time. The ability to use it in the projected Fokker-Planck equation, make it possible to use it in the variational smoothing algorithm of chapter (3.3), or at least the variational smoothing algorithm with projected constraints in section 3.6. The skew-normal density leads naturally to a linear drift function with an additive skewness potential, and reduces to the linear case as the skewness of the drift lends to zero. The first moment equation in the skew-normal variational smoother is derived in section 4.3.3.1 and future work will look at completing these derivations.

Chapter 5

Optimal Gaussian smoothing

Contribution

- The variational GP algorithm is considered in the context of other temporal GP machine learning algorithms. The formal relations between variational GP smoothing and GP regression, kernel regression, and assumed Gaussian smoothing are identified. Additionally, improvements to the gradients method used to implement the variational GP method are proposed. These improvements utilise the Hessian of the free energy which has yet to be used in variational GP smoothing.

5.1 Introduction

Independently, both Gaussian mixture models and exponential families have universal approximation abilities to any continuous density function (Mclachlan & Peel, 2000, Wainwright & Jordan, 2008, respectively). The Gaussian density sits at the intersection of both these families; being a Gaussian mixture model with only one component and an exponential family density with only linear and quadratic sufficient statistics. Therefore any progress made with the trusty Gaussian is relevant to both the more general settings. Even with its downsides (e.g. its inability to capture multi-modality) the humble Gaussian is a worthy choice of parametric model for any investigation. All distributions are dominated by their first and second moments, and when learning is based upon a single (partially observed) sample-path, the descriptive power of a Gaussian is in many cases powerful enough to accurately track the position of such a path.

In this chapter, the variational GP algorithm is considered in the context of other GP smoothing algorithms such as GP regression, kernel regression, and assumed Gaussian smoothing. There are some key reasons for doing this. For linear temporal problems, i.e. one dimensional ordered input, kernel methods and simple stationary GP regression do not have the capability to represent the complex a priori dynamics and observation models that can be dealt with easily in a Kalman smoothing framework. Most kernel and GP regression methods do not scale well with the number of observations because of the required inversion of a covariance matrix, though this can be reduced to the same complexity as a Kalman smoother if Markov assumptions are made. While the variational GP algorithm reduces to Kalman smoothing when the dynamics and observation model are linear, neither kernel methods, GP regression, or Kalman smoothing have the capability to deal with nonlinear dynamics or observation models. Kernel methods and GP regression can easily handle higher order noise correlations without having to increase the dimension of the problem, as Kalman and variational GP smoothing methods would. Kernel methods and GP regression can also easily handle high dimensional spatial input. For algorithms analogous to Kalman and variational GP smoothing methods, i.e. under the assumption of continuous and ordered input, this would require the definition and integration of a high dimensional stochastic partial differential equation. This is not the Fokker-Planck equation, but an equation describing the interplay between the input spatial variables and the output. Additionally, kernel methods can handle a range of loss functions beyond squared loss. These types of generalised costs on paths would be required for a more general relation to optimal control. If there is any hope of synergism, then the formal relations between all the methods needs to made explicit. This is done, somewhat, in an finite dimensional setting in Steinke & Schölkopf (2008). Here a similar approach is used, and extended, to discuss the key concepts in the general setting. To make the investigation tractable, and so as not to confuse notation, analysis is restricted to temporal problems.

The second part of the chapter looks at the recently proposed Assumed Gaussian smoothing algorithm. While this algorithm is fast and highly stable, it is restricted to one forward and backward pass of the data. Some deterministic analysis for the GP projection filter on the double well problem is given, and the extension to the incorporation of more general deterministic analysis from dynamical systems theory is discussed. The analysis goes some way to explaining why the GP projection smoother does not well represent the transitions.

5.2 Kernels, covariance operators, and linear models

The connection between regularisation operators and reproducing kernels is well known (Smola et al., 1998), as is the connection between reproducing kernels and Gaussian processes (Bertinet & Agnan, 2004, van der Vaart & van Zanten, 2008). Recently, all the objects discussed in the introduction of this section were considered under a unified framework (Steinke & Schölkopf, 2008) specialised to the finite dimensional setting, i.e. a finite discretisation of time. The following section builds upon this work, making certain extensions to the form of linear state space model considered and trying to work in a more general setting as much as possible. The solution of the variational GP algorithm is a linear state space model describing the evolution of a Gaussian process. The hope is that the connections made will facilitate the transfer of ideas between the variational GP method and standard GP and kernel regression.

5.2.1 Regularisation operators

In Steinke & Schölkopf (2008) the abstract notion of a *covariance operator* is placed at the centre of the unified theory, and all ideas such as GP's and kernels bloom from the definition of the covariance operator like leaves. A different approach is taken here. One of the leaves that comes off the definition of the covariance operator is the notion of a regularisation operator, but the map from covariance operator to regularisation operator is not unique. Thus, if we place the covariance operator at the centre of the theory, there will exist ambiguity about which regularisation operator we are working with. This abstract map can be seen in concrete form if one tries to rebuild a linear state space model from a Gaussian process. The following general exposition is brief. It is given to highlight the relations in the general (possibly infinite dimensional) setting. We can afford this brevity because, in the following section, the variational GP solution is considered in the discrete time setting, where a lot of the technical machinery can be ignored.

Definition 5.1 (Regularisation operator (Scholkopf & Smola, 2001))**.** *A* regularisation operator *$\mathcal{R}$ is defined as a linear map from a dot product space of functions into a dot product space.*

Though this is quite a general definition, regularisation operators will almost always be a linear combination of differential operators, for example

$$\mathcal{R}x = (\partial_t^2 + w^2)x. \tag{5.1}$$

In this example, the equation $\mathcal{R}x = 0$ defines the dynamics, and $x(t)$ the path of an *undamped harmonic oscillator*. More generally, regularisation operators are used to encode properties of an underlying system. Given some data $(y_i, t_i)_{i=1}^m$, it is simple to see how minimising the functional

$$I(x) = \sum_{i=1}^{m} \text{ConvexLoss}(y_i, x(t_i)) + \lambda ||\mathcal{R}x||_{L^2}, \tag{5.2}$$

will prioritise solutions that meet the properties encoded in the equation $\mathcal{R}x = 0$. The set of solutions $\text{Null}(\mathcal{R}) = \{x|\mathcal{R}x = 0\}$ is defined the *null space* of $\mathcal{R}$. It would be nice if $\text{Null}(\mathcal{R})$ contained a unique solution, but generally $\text{Null}(\mathcal{R})$ defines a subspace, e.g.

$$\text{Null}(\partial_t) = \{x|x(t) = \text{constant}, \forall t\}. \tag{5.3}$$

Let $\mathcal{R}^*$ denote the adjoint of $\mathcal{R}$, e.g. $\langle x, \mathcal{R}x' \rangle_{L^2} = \langle \mathcal{R}^* x, x' \rangle_{L^2}$, and let $(\mathcal{R}^*\mathcal{R})^\dagger$ denote the *Moore-Penrose pseudo inverse*.

Definition 5.2 (Covariance operator)**.** *If $(\mathcal{R}^*\mathcal{R})^\dagger$ is linear, continuous, positive, self-adjoint and finite trace, then $(\mathcal{R}^*\mathcal{R})^\dagger$ defines a covariance operator, and is denoted*

$$\mathcal{C} = (\mathcal{R}^*\mathcal{R})^\dagger. \tag{5.4}$$

The above definition is over-cautious. Linearity, positiveness, and self adjointness follow by definition of $\mathcal{R}^*\mathcal{R}$. Continuity and finite trace follow from the pseudo inverse operation, i.e. boundedness. There is no harm in including these properties in the definition because together they define an *abstract* covariance operator, without the need for a regularisation operator. For ease of exposition, let us now specialise to $L^2([0,T])$ with inner product $\langle x, x\rangle_{L^2} = \int_0^T x(t)x'(t)dt$. The covariance operator $\mathcal{C}$ can be interpreted in concrete form through the use of a symmetric positive definite *kernel* $k(s,t)$ such that

$$\mathcal{C}x(\cdot) = \int k(\cdot,t)x(t)dt. \tag{5.5}$$

Let δ_t denote the evaluation operator centered at t such that, for all $x \in L^2$ and $t \in [0,T]$,

$$\langle \delta_t, x\rangle_{L^2} = x(t). \tag{5.6}$$

Then it holds that

$$k(s,t) = \langle \delta_s, \mathcal{C}\delta_t\rangle_{L^2}. \tag{5.7}$$

Using $k(s,t)$ it is possible to define a *Gaussian process* $\{x_t\}_{t\in[0,T]}$ with mean $\mu(t) = \langle x_t\rangle$ and covariance $k(t,s)$ such that

$$\begin{aligned} \langle x_t\rangle &= \sum_i \alpha_i k(t_i,t), \quad \alpha_i \in \mathbb{R}. && (5.8)\\ \big\langle (x_t - \langle x_t\rangle)(x_s - \langle x_s\rangle)\big\rangle &= k(t,s). && (5.9) \end{aligned}$$

Or to define a reproducing kernel Hilbert space $(\mathcal{H},k)$ such that

$$\mathcal{H} = \overline{\{x|x = \sum_i \alpha_i k(t_i,t), \quad \alpha_i \in \mathbb{R}\}} \tag{5.10}$$

where $\overline{S}$ denotes the *completion* of the set S. Importantly for the regularisation problem in (5.2), $(\mathcal{H},k)$ has an inner product define $\langle\cdot,\cdot\rangle_k$ such that

$$\langle x, x'\rangle_k := \langle x, \mathcal{C}^{-1}x'\rangle_{L^2} = \langle \mathcal{R}x, \mathcal{R}x'\rangle_{L^2}, \tag{5.11}$$

and also

$$\begin{aligned} \langle x, k(t,\cdot)\rangle_k &= x(t) && (5.12)\\ ||x||_k &= ||\mathcal{R}x||_{L^2}. && (5.13) \end{aligned}$$

Equation (5.12) is referred to as the *reproducing property*. Equation (5.13) is the fundamental link between kernel methods and regularisation network, and goes some way to explaining why kernel methods work. The completion in equation (5.10) is an abstract operation and, in fact, becomes a space of functions through the reproducing property (van der Vaart & van Zanten, 2008). The above outline hardly touches the surface of the theory underlying regularisation operators $\mathcal{R}$, Gaussian processes $\mathcal{GP}(\mu, k)$, and reproducing kernels $k(s,t)$. The point to take away from this is the relation $\mathcal{R} \to \mathcal{C} \to k$. This is in contrast to the approach in Steinke & Schölkopf (2008) and more like the one in Steinke & Schlkopf (2006). A differential equation generates a regularisation operator, which leads to a covariance operator, which leads to the common positive definite methods of machine learning. In Steinke & Schlkopf (2006) they discuss learning the properties of a differential equation from noisy data through approaches such as *kernel learning* (Argyriou et al., 2005). This seems an *ill posed* problem given the non unique nature of the map from covariance operator back to regularisation operator, i.e. $\mathcal{R} = (\mathcal{C}^{\frac{1}{2}})^{\dagger}$. A better posed problem might be to *learn the kernel* through methods that learn dynamics of a (stochastic) differential operator, such as the Kalman filter or variational methods covered here.

5.2.2 Linear state space models

This section puts the ideas of the previous section into concrete forms. The regularisation operator is defined from a linear state space model. The exposition is made simple through a discretisation of time. While the solutions of the variational GP algorithm are given in continuous time form, they have to discretised in implementation. This ensures that the following ideas are applicable to the variation GP solutions. In contrast to Steinke & Schölkopf (2008), here time varying models are considered. Consider a linear model described by the following equations

$$d\mathbf{x}(t) = \mathbf{A}(t)\mathbf{x}(t)dt + \mathbf{b}(t) + \sqrt{\mathbf{D}}(t)d\mathbf{w}(t), \quad \mathbf{x}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathbf{C}_0). \tag{5.14}$$

It is well known that the mean function $\boldsymbol{\mu}(t)$ and covariance function $\boldsymbol{\Sigma}(t,s)$ of the process described by equation (5.14) can be written in integral form (appendix A.3.6)

$$\begin{aligned} \boldsymbol{\mu}(t) &= e^{\int_0^t \mathbf{A}(s)ds}\mathbf{m}_0 + \int_0^t e^{\int_s^t \mathbf{A}(u)du}\mathbf{b}(s)ds && (5.15) \\ \boldsymbol{\Sigma}(s,t) &= e^{\int_0^s \mathbf{A}(u)du}\mathbf{C}_0 e^{\int_0^t \mathbf{A}^{\mathrm{T}}(u)du} + \int_0^{t\wedge s} e^{\int_u^s \mathbf{A}(v)dv}\mathbf{D}(u)e^{\int_u^t \mathbf{A}^{\mathrm{T}}(v)dv}du. && (5.16) \end{aligned}$$

Note, differentiating $\boldsymbol{\mu}(t)$ and $\boldsymbol{\Sigma}(t,t)$ with respect to t generates the marginal density evolution equations in (3.22) and (3.23). Equations (5.15) and (5.16) will not have a closed form due to the time dependency of the drift and diffusion parameters. Therefore, we require numerical integration to obtain $\boldsymbol{\mu}(t)$ and $\boldsymbol{\Sigma}(t,s)$ restricted to a ascending finite grid of times $\mathcal{T} = \{t_i \in [0,T]\}$. This can be done by defining a discrete time process $\{\mathbf{x}_i\}_{i\in\mathbb{N}_N}$ such that

$$\mathbf{x}_{i+1} = \tilde{\mathbf{A}}(i)\mathbf{x}_i + \tilde{\mathbf{b}}(i) + \sqrt{\tilde{\mathbf{D}}}(i)\boldsymbol{\xi}_i, \quad \mathbf{x}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathbf{C}_0), \tag{5.17}$$

where $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and

$$\begin{aligned}
\tilde{\mathbf{A}}(i) &:= e^{\int_{t_i}^{t_{i+1}} \mathbf{A}(s)ds} && (5.18)\\
\tilde{\mathbf{b}}(i) &:= \int_{t_i}^{t_{i+1}} e^{\int_s^{t_{i+1}} \mathbf{A}(u)du}\mathbf{b}(s)ds && (5.19)\\
\tilde{\mathbf{D}}(i) &:= \int_{t_i}^{t_{i+1}} \Big(e^{\int_s^{t_{i+1}} \mathbf{A}(u)du}\Big)\mathbf{D}(s)\Big(e^{\int_s^{t_{i+1}} \mathbf{A}(u)du}\Big)^{\mathrm{T}} ds. && (5.20)
\end{aligned}$$

It can be shown that the mean and covariance of $\mathbf{x}_i$ matches $\mathbf{x}_t$ on the grid $\mathcal{T}$. Define $\mathbf{A}(i:j) = \mathbf{A}(i)\mathbf{A}(i-1)\cdots\mathbf{A}(j+1)\mathbf{A}(j)$ for $i \geq j$, and note $(\mathbf{A}(i:j))^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}(j:i)$. Then the discrete time mean $\boldsymbol{\mu}(i)$ and covariance $\boldsymbol{\Sigma}(i,j)$ can be written

$$\begin{aligned}
\boldsymbol{\mu}(i) &= \tilde{\mathbf{A}}(i:0)\mathbf{m}_0 + \sum_{l=1}^{i} \tilde{\mathbf{A}}(i:l)\tilde{\mathbf{b}}(l) && (5.21)\\
\boldsymbol{\Sigma}(i,j) &= \tilde{\mathbf{A}}(i:0)\mathbf{C}_0\tilde{\mathbf{A}}^{\mathrm{T}}(0:j) + \sum_{l=1}^{i\wedge j} \tilde{\mathbf{A}}(i:l)\tilde{\mathbf{D}}(l)\tilde{\mathbf{A}}^{\mathrm{T}}(l:i). && (5.22)
\end{aligned}$$

Alternatively, now that we have discretised the system we can deal with it as a finite dimensional Gaussian process. For $\mathbf{X} = \mathrm{vec}(\mathbf{x}_0,\ldots,\mathbf{x}_{N-1})$, equation (5.17) can be written

$$\boldsymbol{\xi} = \tilde{\mathcal{R}}\mathbf{X} - \mathbf{u} \tag{5.23}$$

where $\boldsymbol{\xi} = \mathrm{vec}(\boldsymbol{\xi}_0', \boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_{N-1})$, $\mathbf{u} = \mathrm{vec}\Big(\mathbf{C}_0^{-\frac{1}{2}}\mathbf{m}_0, \tilde{\mathbf{D}}_0^{-\frac{1}{2}}\tilde{\mathbf{b}}_0, \ldots, \tilde{\mathbf{D}}_{N-1}^{-\frac{1}{2}}\tilde{\mathbf{b}}_{N-1}\Big)$, and $\tilde{\mathcal{R}}$ denotes the difference operator

$$\tilde{\mathcal{R}} = \begin{pmatrix} \mathbf{C}_0^{-\frac{1}{2}} & & & \\ -\tilde{\mathbf{D}}_0^{-\frac{1}{2}}\tilde{\mathbf{A}}_0 & \tilde{\mathbf{D}}_0^{-\frac{1}{2}} & & \\ & \cdots & \cdots & \\ & & -\tilde{\mathbf{D}}_{N-1}^{-\frac{1}{2}}\tilde{\mathbf{A}}_{N-1} & \tilde{\mathbf{D}}_{N-1}^{-\frac{1}{2}} \end{pmatrix}. \tag{5.24}$$

Then, it can be easily checked from equations (5.21) and (5.22) that

$$\begin{aligned}
\mathrm{vec}(\boldsymbol{\mu}) &= \tilde{\mathcal{R}}^{-1}\mathbf{u} && (5.25)\\
\boldsymbol{\Sigma}(i,j) &= (\tilde{\mathcal{R}}^{\mathrm{T}}\tilde{\mathcal{R}})^{-1}_{block[i,j]}. && (5.26)
\end{aligned}$$

in analogy to definition 5.2. So from equation (5.14) restricted to a finite grid $\mathcal{T}$, we have constructed a Gaussian process in equations (5.21) and (5.22), and a regularisation operator in equation (5.24), and we know the two relate through equations (5.25) and (5.26). All that is left is for us to build a function space. This is relatively simple in the discrete time setting. First, let us define the covariance operator

$$\tilde{\mathcal{C}} = (\tilde{\mathcal{R}}^{\mathrm{T}}\tilde{\mathcal{R}})^{-1}. \tag{5.27}$$

Now let us define the vectors $\boldsymbol{\delta}_i = \mathbf{e}_i \otimes \mathbf{I}_{d_\mathbf{x}}$ such that, for any block matrix $\mathbf{B}$,

$$\boldsymbol{\delta}_i^{\mathrm{T}}\mathbf{B} = \mathbf{B}_{block[i,:]}, \quad \mathbf{B}\boldsymbol{\delta}_j = \mathbf{B}_{block[:,j]}, \quad \boldsymbol{\delta}_i^{\mathrm{T}}\mathbf{B}\boldsymbol{\delta}_j = \mathbf{B}_{block[i,j]}. \tag{5.28}$$

The objects $\boldsymbol{\delta}_i$ act as evaluation maps for finite dimensional spaces of vector valued functions, in analogy to the evaluation maps δ_t. Let us define a positive definite matrix valued kernel function $\mathbf{K}(i,j) = \boldsymbol{\Sigma}(i,j)$ on $\mathcal{T}$. It is simple to see from equation (5.26) that

$$\mathbf{K}(i,\cdot) = \boldsymbol{\delta}_i^{\mathrm{T}}\tilde{\mathcal{C}}, \quad \mathbf{K}(\cdot,i) = \tilde{\mathcal{C}}\boldsymbol{\delta}_i, \quad \mathbf{K}(i,j) = \boldsymbol{\delta}_i^{\mathrm{T}}\tilde{\mathcal{C}}\boldsymbol{\delta}_j, \tag{5.29}$$

in analogy to equation (5.7). Due to dealing with finite dimensional paths $\mathbf{X} \in \mathbb{R}^{N\times d_\mathbf{x}}$, any innerproduct on the space of paths $\mathbb{R}^{N\times d_\mathbf{x}}$ defines an RKHS. Thus, we simply define $\langle\cdot,\cdot\rangle_\mathbf{K}$ such that

$$\langle\mathbf{X},\mathbf{X}'\rangle_\mathbf{K} = \mathbf{X}^{\mathrm{T}}\tilde{\mathcal{C}}^{-1}\mathbf{X} = \mathbf{X}^{\mathrm{T}}\tilde{\mathcal{R}}^{\mathrm{T}}\tilde{\mathcal{R}}\mathbf{X} = ||\tilde{\mathcal{R}}\mathbf{X}||^2, \tag{5.30}$$

in analogy to equation (5.11). From the relation $\boldsymbol{\Sigma}(i,j) = \mathbf{K}(i,j)$ and the identities in (5.29), it holds that

$$\begin{aligned} \langle\mathbf{X},\mathbf{K}(\cdot,i)\rangle_\mathbf{K} &= \mathbf{X}^{\mathrm{T}}\tilde{\mathcal{C}}^{-1}\mathbf{K}(\cdot,i) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\delta}_i = \mathbf{x}_i & (5.31)\\ \langle\mathbf{K}(\cdot,i),\mathbf{X}\rangle_\mathbf{K} &= \mathbf{K}(i,\cdot)\tilde{\mathcal{C}}^{-1}\mathbf{X} = \boldsymbol{\delta}_i^{\mathrm{T}}\mathbf{X} = \mathbf{x}_i^{\mathrm{T}}, & (5.32)\end{aligned}$$

proving the reproducing property of $\mathbf{K}$. So from the the linear state space model in equation (5.14), we have defined a discrete time regularisation operator $\tilde{\mathcal{R}}$, a Gaussian process $\mathcal{GP}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, and reproducing kernel $\mathbf{K}$. Interesting future directions would be to look at the possibility of characterising the complexity of the solutions of the variational GP algorithm, using kernel approaches for studying function class complexity. Note that this would require additional material to deal with the non stationary time varying nature of the kernel $\mathbf{K}(s,t)$. In the other direction, there is the important point that the kernel $\mathbf{K}(s,t)$ encodes properties of the prior non linear model. In the variational GP algorithm, the data has already been assimilated, so $\mathbf{K}(s,t)$ cannot be used as a standard prior kernel. The use of non stationary kernel and covariance function in machine learning is rare, and the use of nonlinear differential operators is non existent due to the break down of the theory in section 5.2.1, most importantly that $||\mathcal{R}x||$ would define a norm in the general setting. Therefore, the variational GP algorithm and the kernel and covariance interpretations above, suggest a possible, though at present computationally intensive, way to incorporate non stationarity and nonlinearity into standard kernel methods.

5.2.3 Stationary linear model

In certain cases, zero-mean GPs in infinite dimensional spaces can be converted into explicit state-space forms. Some of the content here is taken from Hartikainen & Särkkä (2010). The most common types of covariance functions used in Gaussian processes learning include the class of *Matérn* kernels and the *squared exponential* kernel. The Matérn kernel is given by

$$k^{(\sigma,l,\nu)}_{\text{Matern}}(\tau) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \bigg(\frac{\sqrt{2\nu}}{l} \tau \bigg)^{\nu} K_\nu \bigg(\frac{\sqrt{2\nu}}{l} \tau \bigg), \tag{5.33}$$

where K_ν is a modified Bessel function of second kind Abramowitz & Stegun (1964), Γ denotes the Gamma function, and (σ, l, ν) are magnitude, length, and smoothness parameters, respectively. Let us define $\lambda = \frac{\sqrt{2\nu}}{l} > 0$, and let $D(\sigma, \lambda, \nu)$ denote the spectral density

$$D(\sigma, \lambda, \nu) = \frac{2\sigma^2 \sqrt{\pi} \lambda^{2\nu} \Gamma(\nu + \frac{1}{2})}{\Gamma(\nu)}. \tag{5.34}$$

For $\nu = \frac{1}{2}$, the GP $\mathcal{GP}(0, k^{(\sigma,\lambda,\frac{1}{2})}_{\text{Matern}})$, which will be referred to as a *Matérn-1 GP*, is generated according to the steady-state of the SDE

$$dx_t = -\lambda x_t dt + \sqrt{D(\sigma, \lambda, \tfrac{1}{2})} dw_t. \tag{5.35}$$

For $\nu = \frac{3}{2}$, the GP $\mathcal{GP}(0, k^{(\sigma,\lambda,\frac{3}{2})}_{\text{Matern}})$, which will be referred to as a *Matérn-3 GP*, is generated according to the steady-state of the SDE

$$d\mathbf{x}_t = \begin{pmatrix} -2\lambda & -\lambda^2 \\ 1 & 0 \end{pmatrix} \mathbf{x}_t dt + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \sqrt{D(\sigma, \lambda, \tfrac{3}{2})} dw_t, \tag{5.36}$$

where $\mathbf{x}_t = (x^{(0)}_t, x^{(1)}_t)$ is a two dimensional state-vector. For $\nu = \frac{5}{2}$, the GP $\mathcal{GP}(0, k^{(\sigma,\lambda,\frac{5}{2})}_{\text{Matern}})$, which will be referred to as a *Matérn-5 GP*, is generated according to the steady-state of the SDE

$$d\mathbf{x}_t = \begin{pmatrix} -3\lambda & -3\lambda^2 & -\lambda^3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \mathbf{x}_t dt + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \sqrt{D(\sigma, \lambda, \tfrac{5}{2})} dw_t, \tag{5.37}$$

where $\mathbf{x}_t = (x^{(0)}_t, x^{(1)}_t, x^{(2)}_t)$ is a three dimensional state-vector. New dimensions need to be introduced to represent the higher-order noise correlations that follow from *smoother* covariance functions, i.e. $\nu > \frac{1}{2}$. The primary direction $x^{(1)}$ describes the position of paths. The new dimensions can be thought of as *generalised coordinates*; controlling position, velocity, and higher-order motions $(x, x', x'', \cdots, x^{(k)})$ (Friston et al., 2008). As the smoothness parameter is taken to infinity, $\nu \to \infty$, we obtain the squared exponential kernel,

$$k^{(\sigma,l)}_{\text{SE}}(\tau) = \sigma^2 \exp\bigg(-\frac{\tau^2}{2l^2} \bigg). \tag{5.38}$$

A process with covariance function $k^{(\sigma,l)}_{\text{SE}}$ is infinitely differentiable, meaning it cannot be represented by a finite-dimensional Markov process. This can be remedied by taking an arbitrarily large finite-dimensional approximation, where it was shown in Hartikainen & Särkkä (2010) that dimension six appears to give reasonable approximation errors. The above exposition has skipped the definition of abstract regularisation operators and covariance operators, and has moved directly from covariance function to state space model. While the operations in this section can be reversed to obtain the corresponding covariance function from the given Markov process, the available state-space models that lead to closed-form covariance functions are few and far between, when compared to the range of general Markov processes covered in the previous chapters. It is not known how easily the above methods can be extended to multidimensional processes, therefore it is assumed the above methods are restricted to only one dimension. The above methods constrain the linear dynamics and diffusions considerably and are also only applicable to time-invariant systems. To generate the corresponding RKHS, let us restrict attention to the Matérn-1 GP. Consider the SDE in equation (5.35) with $\lambda > 0$ and $D(\sigma, \lambda, \frac{1}{2}) = D > 0$ for some constant D. The corresponding process is (with rescaling) a Matérn-1 GP $\mathcal{GP}(0, k)$ with covariance function $k(s,t) = \frac{D}{2\lambda}\exp(-\lambda|s-t|)$ (Rasmussen & Williams, 2005). If the noise fluctuations are removed from equation (5.35) then we can define a differential $x' = \frac{dx}{dt}$ of x, and a space

$$\mathcal{X} = \left\{ x \in L^2 \middle| ||x' + \lambda x||_{L^2}^2 + 2\lambda|x(0)|^2 < \infty \right\} \tag{5.39}$$

where $||\cdot||_{L^2}$ is the norm in $L^2([0,T],\mathbb{R})$. The constraint on the initial condition has been introduced to ensure all paths begin in $\mathbb{R}$, but apart from this no other assumptions are made on $x(0)$. It is possible to show that $\mathcal{X}$ is a Reproducing kernel Hilbert space (RKHS) with reproducing kernel

$$k(s,t) = \frac{\exp\big(-\lambda|s-t|\big)}{2\lambda}. \tag{5.40}$$

Though this is a standard result (see Parzen (1961), Kailath (1971), and Bertinet & Agnan (2004)), an exact proof could not be found and one is provided in appendix A.3.7. Equation (5.40) is often referred to as the *Laplace* kernel, and is equivalent (with rescaling) to the covariance function of the Matérn-1 GP $\mathcal{GP}(0, k)$. The above formulation can be extended to higher-order differential operators, i.e. $||\sum_i \lambda_i \partial_t^{(i)} x||_{L^2}^2$, to obtain analogous versions of the general Matérn kernels in section 5.2.3 (see Bertinet & Agnan (2004)). While we benefit a lot from using a *compact*[1] input space $[0,T]$, it prevents us from using the *Fourier analysis* approach of (Rasmussen & Williams, 2005, section 6.2.1) to move between kernels and differential operators.

In summary, explicit connections between state space models, covariance functions, and reproducing kernels can only be done in continuous time for a few specific state space models. This prevents an explicit interpretation, in terms of GPs and function spaces, of the general solutions of the variational GP algorithm in continuous time form. While this doesn't pose too much of a problem given the discretisation method of the previous section, it is desirable to have results that do not depend on the graining of the discretisation. One possible future direction,

[1] i.e. *closed* and *bounded* (Kolmogorov & Fomin, 1975).

is the idea of integrating the noise correlation capturing abilities of the above kernels into a variational formulation. To integrate higher level noise correlations into the state space model requires an increase in dimension, as discussed in section 4.1.3. It might be possible to use the variational GP approximation at one level, and then collect all the higher level correlations into an upper level and apply GP regression with a higher order Matérn kernel. At present though, this is pure speculation. Another future investigation be would to formulate the above relations in algorithmic terms. Kalman smoothing methods avoid explicit matrix inversion and therefore are computationally superior for linear time invariant state space models. But to incorporate higher order noise correlations requires integer multiplications of the dimension of the problem. Thus there must be a point at which noise correlations become too much, and superiority returns to kernel based methods.

5.3 Assumed Gaussian smoothing

The previous section dealt with relations between some important properties of the model. While it was suggested that the analysis in section 5.2.2 could be applied to the variational GP approximation, i.e. post learning, inference was not considered per say. This section considers the two main alternative methods to the variational GP algorithm for providing GP posterior solutions. The first approach in section (5.3.1) considers general GP regression as is seen in machine learning. This method relies upon the general rule for conditioning a Gaussian measure. While general covariance functions can be used, the method is a more general version of the one encountered in Rasmussen & Williams (2005). The method is phrased in the language of vector valued functions, is restricted to temporal input (though almost all the workings hold for more general inputs), and can handle general linear observation operators. Despite these differences, it should still be considered the standard GP regression method of machine learning when compared to other methods. The second approach is a continuous time smoothing algorithm that subsumes many approximate smoothing methods. It is an assumed density smoothing algorithm with Gaussian assumption, but a continuous time formulation only recently derived (Särkkä & Sarmavouri, 2011). It subsumes all related extended and unscented continuous time smoothing methods Särkkä (2010). Importantly, it is very fast and stable in experiment.

5.3.1 Temporal GP regression

Temporal Gaussian process models are characterised by a mean function $\boldsymbol{\mu}(\cdot)$ and covariance function $\boldsymbol{\Sigma}(\cdot,\cdot)$, and subsume the class of linear state-space models. Importantly, while they can represent higher-order noise fluctuations than the ones generated by the Wiener process in (2.21), they are not able to represent state-space models with nonlinear drifts. All GP inference models can be phrased in the language of the Gaussian conditioning theorem A.2, and solved using the conditioned Gaussian equations in (A.18). Let us restrict attention to the case of temporal GP models over paths $\mathbf{x}(\cdot)$, with inputs $t \in [0, T]$ and output $\mathbb{R}^{d_\mathbf{x}}$. Let us assume we have a prior Gaussian process $\mathcal{GP}_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ over paths $\mathbf{x}(t)$, with vector valued mean $\boldsymbol{\mu}_0(t)$ and matrix-valued covariance function $\boldsymbol{\Sigma}_0(s, t)$. Let us assume we have a linear observation model

$$\mathbf{y}_i = \mathbf{H}(t_i)\mathbf{x}(t_i) + \boldsymbol{\nu}(t_i) + \boldsymbol{\eta}_i, \quad i = 1, \ldots, m, \tag{5.41}$$

where $\mathbf{H}(t) \in \mathbb{R}^{d_\mathbf{y} \times d_\mathbf{x}}$ and $\boldsymbol{\nu}(t) \in \mathbb{R}^{d_\mathbf{y}}$ are continuous-time functions, and $\boldsymbol{\eta}_i$ is an $\mathbb{R}^{d_\mathbf{y}}$-valued Gaussian random-variable with mean vector zero and variance matrix $\mathbf{R}_i$. We can write the observation-model generating the full dataset $\mathbf{Y} = \text{vec}(\mathbf{y}_1, \ldots, \mathbf{y}_m)$ in the form, corresponding to equation (2.2),

$$\mathbf{Y} = \sum_{i=1}^{m} \langle \delta_{t_i}, \mathbf{H}\mathbf{x} + \boldsymbol{\nu} \rangle_{L^2} \otimes \mathbf{e}_i + \boldsymbol{\eta}, \tag{5.42}$$

where $\langle \cdot, \cdot \rangle_{L^2}$ is the inner product in $L^2([0, T]; \mathbb{R}^{d_\mathbf{y}})$, $\delta_s(\cdot)$ is the evaluation mapping centered at s, $\mathbf{e}_i$ is the i^{th} standard basis in $\mathbb{R}^{d_\mathbf{y}}$, $\otimes$ denotes the Kronecker product, and $\boldsymbol{\eta}$ is an $\mathbb{R}^{m \times d_\mathbf{y}}$-valued Gaussian random-variable with mean-vector zero and block-diagonal variance-matrix $\mathbf{R}$, such that $(\mathbf{R})_{block[i,i]} = \mathbf{R}_i$. In a strict sense, the delta function δ_s is not part of L^2. However, the notation is widely used. It is simple to show that $\mathbf{Y}$ has mean $\boldsymbol{\mu}_\mathbf{Y} \in \mathbb{R}^{d_\mathbf{Y}}$ and covariance $\boldsymbol{\Sigma}_{\mathbf{YY}} \in \mathbb{R}^{d_\mathbf{Y} \times d_\mathbf{Y}}$ given by

$$\boldsymbol{\mu}_\mathbf{Y} = \sum_{i=1}^{m} \mathbf{H}(t_i)\boldsymbol{\mu}_0(t_i) + \boldsymbol{\nu}(t_i) \otimes \mathbf{e}_i \tag{5.43}$$

$$\boldsymbol{\Sigma}_{\mathbf{YY}} = \sum_{i,j=1}^{m} \mathbf{H}(t_i)\boldsymbol{\Sigma}_0(t_i, t_j)\mathbf{H}^{\mathrm{T}}(t_j) \otimes \mathbf{e}_i\mathbf{e}_j^{\mathrm{T}} + \mathbf{R}, \tag{5.44}$$

or

$$\langle \mathbf{y}_i \rangle = \mathbf{H}(t_i)\boldsymbol{\mu}_0(t_i) + \boldsymbol{\nu}(t_i) \tag{5.45}$$

$$\langle \mathbf{y}_i\mathbf{y}_j^{\mathrm{T}} \rangle - \langle \mathbf{y}_i \rangle \langle \mathbf{y}_j \rangle^{\mathrm{T}} = \mathbf{H}(t_i)\boldsymbol{\Sigma}_0(t_i, t_j)\mathbf{H}^{\mathrm{T}}(t_j) + \mathbf{R}_i\delta_{ij}, \tag{5.46}$$

where δ_{ij} denotes the Kronecker Delta. The cross-covariance $\boldsymbol{\Sigma}_{\mathbf{x}(\cdot)\mathbf{Y}}$ is given by

$$\boldsymbol{\Sigma}_{\mathbf{x}(\cdot)\mathbf{Y}} = \sum_{i=1}^{m} \boldsymbol{\Sigma}_0(\cdot, t_i)\mathbf{H}^{\mathrm{T}}(t_i) \otimes \mathbf{e}_i^{\mathrm{T}}. \tag{5.47}$$

Inserting these into theorem A.2, leads to a posterior Gaussian process $\mathcal{GP}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ mean $\boldsymbol{\mu}(t)$ and covariance $\boldsymbol{\Sigma}(s, t)$ given by

$$\boldsymbol{\mu}(t) = \boldsymbol{\mu}_0(t) + \boldsymbol{\Sigma}_{\mathbf{x}(t)\mathbf{Y}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y}) \tag{5.48}$$

$$\boldsymbol{\Sigma}(s, t) = \boldsymbol{\Sigma}_0(s, t) - \boldsymbol{\Sigma}_{\mathbf{x}(s)\mathbf{Y}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}\boldsymbol{\Sigma}_{\mathbf{Y}\mathbf{x}(t)}. \tag{5.49}$$

As a specialisation of the above, consider the *GP* prior $\mathcal{GP}_0(0,k)$ with zero mean and covariance function $k(s,t)$. Also, define $d_{\mathbf{x}} = d_{\mathbf{y}} = 1$, $\mathbf{H}(\cdot) = 1$, $\boldsymbol{\nu}(\cdot) = 0$, and $\mathbf{R}_i = r > 0$. Then it holds that $\boldsymbol{\mu}_{\mathbf{Y}} = 0$, and $\boldsymbol{\Sigma}_{\mathbf{YY}} = \mathbf{K} + r\mathbf{I}_m$, where $\mathbf{K}$ is the kernel matrix $(\mathbf{K})_{ij} = k(t_i,t_j)$, and $\mathbf{I}_m$ is the identity matrix in $\mathbb{R}^{m\times m}$. The cross-covariance becomes

$$\boldsymbol{\Sigma}_{\mathbf{x}(\cdot)\mathbf{Y}} = \sum_{i=1}^{m} k(\cdot,t_i)\otimes \mathbf{e}_i^{\mathrm{T}} =: \mathbf{k}_{\mathcal{T}}(\cdot). \tag{5.50}$$

The posterior mean and covariances in equations (5.48) and (5.49) then become

$$\mu(t) = \mathbf{k}_{\mathcal{T}}(t)\big(\mathbf{K} + r\mathbf{I}_m\big)^{-1}\mathbf{Y} \tag{5.51}$$

$$\Sigma(s,t) = k(s,t) - \mathbf{k}_{\mathcal{T}}(s)\big(\mathbf{K} + r\mathbf{I}_m\big)^{-1}\mathbf{k}_{\mathcal{T}}^{\mathrm{T}}(t), \tag{5.52}$$

which together describe the standard univariate GP regression model (Rasmussen & Williams, 2005). The same formulation is easily done for the multidimensional case $d_{\mathbf{x}} = d_{\mathbf{y}} > 1$, (Alvarez et al., 2011). For this case, the Matérn covariance functions in section 5.2.3 can be used to construct the prior $\mathcal{GP}_0(0, k_{\mathrm{Matern}}^{(\sigma,l,\nu)})$. The problem with this general approach is the requirement to invert the matrix $\boldsymbol{\Sigma}_{\mathbf{YY}}$. If a Matérn kernel is used, then the inference problem could be phrased as a Kalman smoothing problem using the prior state space models in section (5.2.3), which can be solved without computing $\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}$ explicitly. This is not obvious in the above formulation. We see from section 5.2.3 that each kernel in this setting must choose a drift parameter λ and stick to it for all time $t \in [0,T]$. The method has no concept of the prior dynamics in the variational model.

5.3.2 Gaussian projection-smoother

With the simplicity and speed of implementation of the GP projection filter in section 3.4.4.1, it would be desirable to obtain a smoothing backward pass that retained some of the qualities of the projection method. This can in fact be done, again by taking the continuous time limit of a discrete time assumed density smoothing algorithm with Gaussian assumption. It leads to a very fast and stable algorithm. Assume the Gaussian projection filter $p(\mathbf{x},\gamma_t)$ of section (3.4.4.1) has already been computed and let $\boldsymbol{\mu}(t)$ and $\boldsymbol{\Sigma}(t)$ denote its mean and covariance. Let $\bar{\boldsymbol{\mu}}(t)$ and $\bar{\boldsymbol{\Sigma}}(t)$ denote the mean and covariance of a Gaussian smoothing density $p(\mathbf{x},\bar{\gamma}_t)$, such that $\bar{\gamma}_t = \mathrm{vec}(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\Sigma}})$ solve *(backwards in time)*

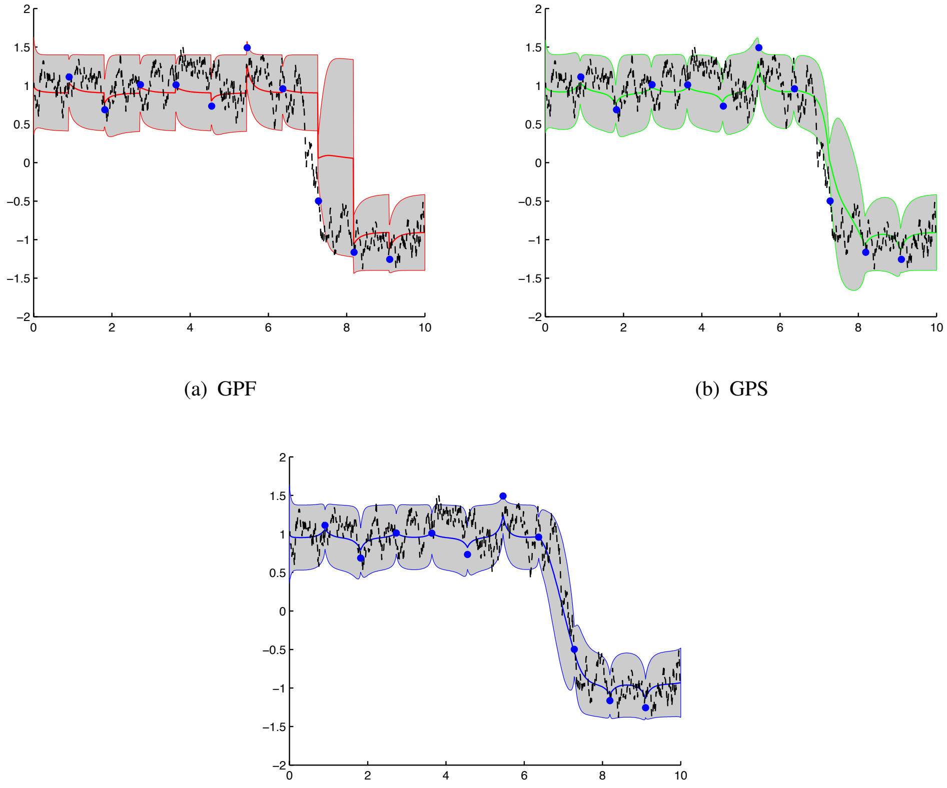
$$\frac{\partial \bar{\boldsymbol{\mu}}_t}{\partial t} = \big\langle \mathbf{f}(\mathbf{x},t)\big\rangle_{p(\mathbf{x},\gamma_t)} + \big\langle \mathbf{f}(\mathbf{x},t)(\mathbf{x}-\boldsymbol{\mu}_t)^{\mathrm{T}} + \mathbf{D}^2(\mathbf{x},t)\big\rangle_{p(\mathbf{x},\gamma_t)}\boldsymbol{\Sigma}_t^{-1}(\bar{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t) \tag{5.53}$$

$$\begin{aligned}\frac{\partial \bar{\boldsymbol{\Sigma}}_t}{\partial t} &= \big\langle \mathbf{f}(\mathbf{x},t)(\mathbf{x}-\boldsymbol{\mu}_t)^{\mathrm{T}} + \mathbf{D}(\mathbf{x},t)\big\rangle_{p(\mathbf{x},\gamma_t)}\boldsymbol{\Sigma}_t^{-1}\bar{\boldsymbol{\Sigma}}_t \\ &\quad + \bar{\boldsymbol{\Sigma}}_t\boldsymbol{\Sigma}_t^{-1}\big\langle(\mathbf{x}-\boldsymbol{\mu}_t)\mathbf{f}^{\mathrm{T}}(\mathbf{x},t) + \mathbf{D}(\mathbf{x},t)\big\rangle_{p(\mathbf{x},\gamma_t)} - \big\langle \mathbf{D}(\mathbf{x},t)\big\rangle_{p(\mathbf{x},\gamma_t)}.\end{aligned} \tag{5.54}$$

This continuous-time smoother algorithm was originally formulated in Särkkä & Sarmavouri (2011). An important point to note is how the expectations are w.r.t. the filter density $p(\mathbf{x}, \boldsymbol{\gamma}_t)$ and *not* the smoothing density $p(\mathbf{x}, \bar{\boldsymbol{\gamma}}_t)$ as might be expected. It turns out that this property leads to very good stability properties. In contrast, smoothing algorithms incorporating expectations over a smoothing density derived from the exact Kolmogorov equations (Leondes et al., 1970) exhibit highly unstable solutions (Särkkä & Sarmavouri, 2011). The method can incorporate the prior dynamics, but is still restricted to one forward and backward pass. An interesting generalisation would be to find a continuous time expectation propagation algorithm. This idea is not that radical an idea, given how the projection filter is essentially the continuous time limit of an assumed density filtering algorithm, and assumed density filtering seeded expectation propagation (Minka, 2001). This topic is discussed in more detail in section 6.1.3.

5.3.3 Unimodal forward-backward approximations

This section provides evidence why Gaussian approximations with only one forward and backward pass of the data are doomed to failure in highly non linear models with sparse observation times. This in turn fuels the argument that multiple iterations such as in the variational GP approximation, or EP-smoothing approximations are superior methods. The Assumed Gaussian smoother of the previous section subsumes, extended, unscented, and other cubature or quadrature based continuous time assumed Gaussian smoothing methods (Särkkä & Sarmavouri, 2011). In this section the exact expectations are computed for the double well problem with identity observation operator. This removes the need for additional numerical methods, and therefore the results here apply to all the assumed Gaussian smoothing methods that use the previously mentioned numerical approximations for computing Gaussian expectations. In figure 5.1 the Gaussian projection filter (GPF) of section 3.4.4.1, the Gaussian projection smoother (GPS) of section 5.3.2, and the variational GP smoother (VGPS) of section (3.3.2) are all applied to the double well problem with a single transition between the wells. It is easily observed how the GPF algorithm is not able to capture the transition. The GP filter subsumes extended Kalman filtering (EKF), and this same result was seen using EKF on the double well problem in (Eyink et al., 2004, Miller et al., 1999). What the results here show, is that this inability to capture the transition is not down the linearisation approximation method used in EKF, it is due to the simple one sweep form of the algorithm. In the plot it can be seen that the GPS algorithm better captures the transition than the GPF due to its reconsideration of the data. It is also clear that the VGPS algorithm captures the transition a lot better than the both GPS and GPF. This is due to the repeated passes of the variational algorithm as it minimises the free energy. To extend the GPS smoother to repeated passes, as opposed to its simple one forward and one backward pass, then a suitable energy objective needs to be defined.



(a) GPF

(b) GPS

(c) VGPS

FIGURE 5.1: Caption

5.3.3.1 Double-well Filter Transitions

The reason for the GPF not picking up the transition can be explained using deterministic analysis. Inserting $f(x) = 4x(\theta - x^2)$ a $D(x,t) = D$ into (3.45) and (3.46) gives a pair of coupled nonlinear differential equations

$$\begin{aligned} \dot{\mu} &= 4\mu(\theta - \mu^2 - 3\sigma) && (5.55) \\ \dot{\sigma} &= 8(\theta - 3(\sigma + \mu^2))\sigma + D && (5.56) \end{aligned}$$

with corresponding Jacobian

$$J^{GP}_{filter}(\mu, \sigma) = \begin{bmatrix} 4(\theta - 3(\mu + \sigma)) & -12\mu \\ -48\mu\sigma & 8(\theta - 3(2\sigma + \mu^2)) \end{bmatrix}. \qquad (5.57)$$

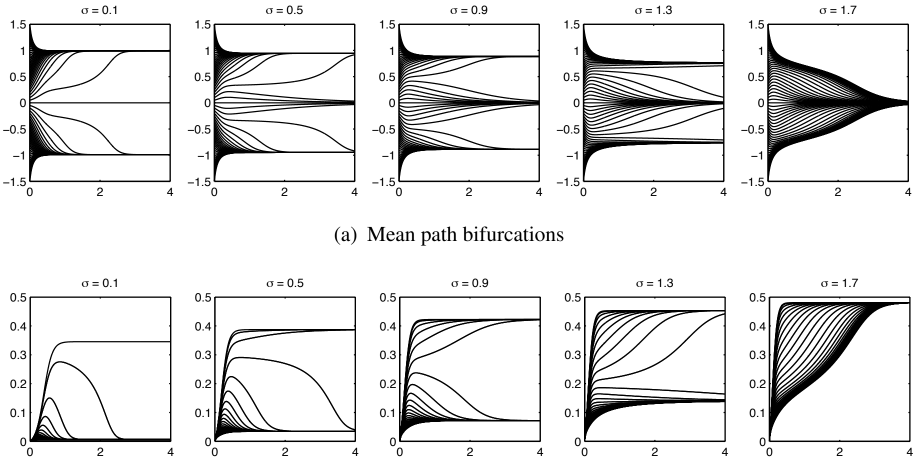
Necessary conditions for (5.55) and (5.56) to be in equilibrium are given by

$$\begin{aligned} \mu &= 0, \pm\sqrt{\theta - 3\sigma} && (5.58) \\ \sigma &= \frac{(\theta - 3\mu^2) \pm \sqrt{(3\mu^2 - \theta)^2 + \frac{3D}{2}}}{6}. && (5.59) \end{aligned}$$

Inserting $\mu = 0$ into (5.59) gives $\sigma' := \frac{\theta+\sqrt{\theta^2+\frac{3D}{2}}}{6}$ (using $\sigma \geq 0$), and the Jacobian at this point is given by

$$J^{GP}_{filter}(0,\sigma') = \begin{bmatrix} 2(\theta - \sqrt{\theta^2 + \frac{3D}{2}}) & 0 \\ 0 & -8\sqrt{\theta^2 + \frac{3D}{2}} \end{bmatrix}. \tag{5.60}$$

Therefore $(\mu, \sigma) = (0, \sigma')$ is a stable equilibrium for all time and all $\theta, D > 0$. Inserting $\mu = \pm\sqrt{\theta - 3\sigma}$ into (5.59) gives $\sigma'' := \theta \pm \sqrt{\theta^2 - \frac{3D}{4}}$. Thus $(\pm\sqrt{\theta - 3s}, \sigma'')$ is only an equilibrium for $D < \frac{4\theta^2}{3}$. The above analysis can be seen visually in figures 5.2. For low diffusion rates, the mean path is very sensitive to initial conditions around zero. When compared to figure 5.1, the inability of GPF to capture the transition makes perfect sense. The mean path of GPF is simply following the underlying dynamics of the prior. Though softened, this problem still permeates through to the smoothing pass. This validates the argument that deterministic smoothing methods that rely upon the simple forward-backward algorithm of exact inference, are dysfunctional. Forward and backward messages interact, just as in discrete time expectation propagation (Heskes & Zoeter, 2002), and multiple passes are required. Note that stochastic methods do not necessarily suffer from the same problem because they are not restricted to a simplified class, such as unimodal densities. The above analysis also suggests some rich pickings for future work. One of the appealing properties of deterministic approximate inference methods in this setting, is the differential equation forms of solutions. This opens up a huge resource of deterministic methods for analysing the solutions. Possible future work would examine which properties of the drift transfer to the variational solution. For polynomial drifts, it is simple to write the solutions in terms of a coupled set of differential equations describing the motion of the mean and covariance parameters over time. For a particular drift, this type



(a) Mean path bifurcations

(b) Covariance path bifurcations

FIGURE 5.2: Double-well mean-plots for projected forward equation for varying diffusion rates.. Each plot contains trajectories starting from initial conditions in the interval $m_0 = -1.5 : 0.05 : 1.5$.

of analysis could be used *a priori* to gauge if a Gaussian approximation is sufficient for the problem.

5.3.4 Free energy comparisons

It is interesting to get an idea of how alternative GP methods perform in a free energy sense. This can be done, at least approximately, by initialising the variational method with alternative methods and observing how solutions behave as a iterations are made. The experiments in this section produce three interesting conclusions. Firstly, it appears that the variational GP method initialised with a highly naive method convergences to the the optimal GP solution just as quickly as if initialised with a more involved method. Secondly, despite the global speed of convergence, more involved initialisations appear to find near optimal solutions in linear regions. For a large amount of the allocated time interval, paths lie within such regions around equilibrium points. This suggests that it might be possible to used suboptimal methods, such as GPS, to find near optimal solutions for the majority of the time interval, and then apply adaptive variational methods just to nonlinear regions, identified by tracing the free energy as a function of time. Thirdly, key to the free-energy formulation is the need for the approximating process to be *absolutely continuous* with respect to the prior process. If the prior is a one dimensional process driven by a Wiener process dw_t, then only the Matérn-1 GP from the above processes (with some rescaling) is absolutely continuous with respect to the prior. For the higher order Matérn GPs, the state-space vector moves in dimensions the prior cannot, and therefore the free energy for any Matérn GP with $\nu > \frac{1}{2}$ will be *infinite*. This property is experienced in the experiments. A simple numerical solution is used to convert mean and covariances into state-space form and reconstruct the corresponding linear drift approximately. While this map is not unique, it is argued that any linear drift that satisfies the evolution of the mean and covariance is sufficient. This allows a variety of GP initialisation methods to be tested. It could also allow for a *hybrid* parameter estimation procedure, where suboptimal GP approximations are inserted into the free-energy in a "plug-and-play" fashion. With only additional few iterations of an adaptive variational GP method, this could lead to improved speed over previous naive initialisations.

5.3.4.1 Converting moment GPs to canonical form

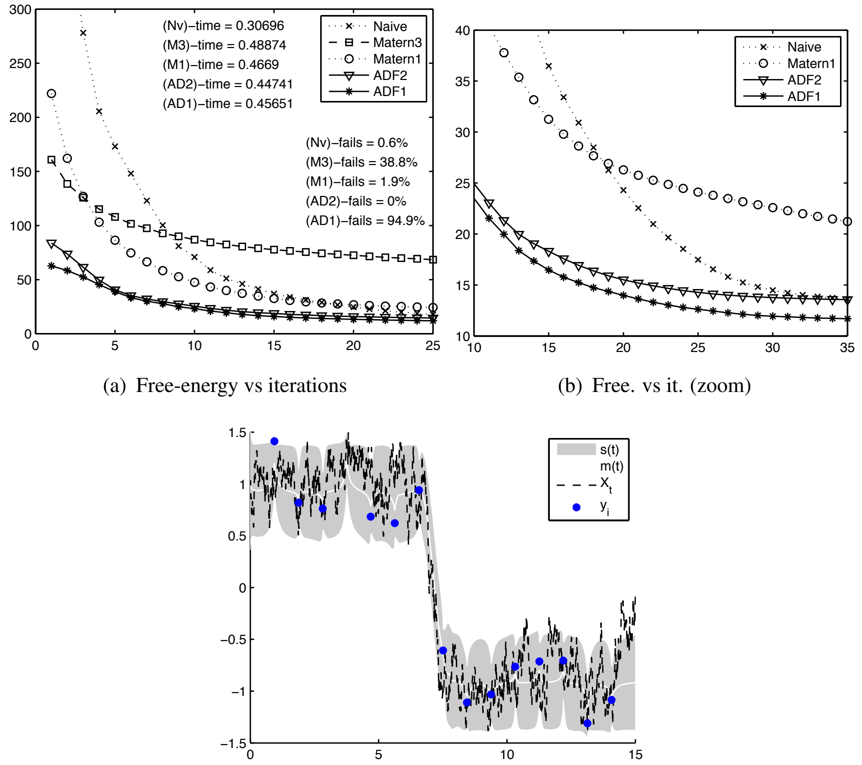
Consider the Gaussian moment equations in (3.22) and (3.23). Equation (3.23) is close to being a *Lyapunov equation*. With a little work, equations (3.22) and (3.23) can be written in terms of $\mathbf{A}(t)$ and $\mathbf{b}(t)$, such that

$$\begin{aligned}\mathbf{b}(t) &= \mathbf{A}(t)\boldsymbol{\mu}(t) - \dot{\boldsymbol{\mu}}(t) && (5.61)\\ \mathrm{vec}\big(\mathbf{A}(t)\big) &= \Big(\big(\boldsymbol{\Sigma}(t)\otimes\mathbf{I}\big) + \mathbf{T}_{mm}\big(\boldsymbol{\Sigma}(t)\otimes\mathbf{I}\big)\Big)^{-1}\mathrm{vec}\Big(\dot{\boldsymbol{\Sigma}}(t) - \mathbf{D}(t)\Big) && (5.62)\end{aligned}$$

where $\mathbf{T}_{mm}$ denotes the matrix such that $\text{vec}(\mathbf{A}) = \mathbf{T}_{mm}\text{vec}(\mathbf{A}^{\text{T}})$ for every possible form of $\mathbf{A}$. The idea is to learn an initial GP $\mathcal{GP}_0(\boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)})$ and insert $\boldsymbol{\mu}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$ into equations (5.61) and (5.62) to obtain initial parameter values $\mathbf{A}^{(0)}$ and $\mathbf{b}^{(0)}$ for use in the variational Gaussian smoothing algorithm. While this may not be possible in the general setting, certain factors need to be remembered. Firstly, any initial GP, $\mathcal{GP}_0$, will have been learned on a discretised grid. If the $\mathcal{GP}_0$ was computed from its own state-space model, as is the case for the assumed Gaussian smoother of section 5.3.2, then $\mathbf{A}^{(0)}$ and $\mathbf{b}^{(0)}$ will already be available in discrete-time form. If the $\mathcal{GP}_0$ was not computed from an explicit state-space form, then it will still be possible to approximate $\dot{\boldsymbol{\mu}}(t)$ and $\dot{\boldsymbol{\Sigma}}(t)$ numerically for all time points on the discretised grid. While, this might seem a bit ad hoc, it needs to be remembered that these initialisations do not need to be exact, they just need to provide the variational smoothing algorithm with starting conditions that are in regions that generate the desired type of GP on the first forward pass. Assuming all the required objects in equations (5.61) and (5.62) are available, then equation (5.62) still requires the inversion of an $m^2 \times m^2$ matrix where m is the dimension of the state space. Naively, this computation is of the order $O(d_{\mathbf{x}}^6)$ which is far is too large for any multidimensional problems. But given the form of the matrix to be inverted it is likely that this can be improved significantly. An alternative is to constrain $\mathbf{A}(t)$ to be symmetric, remembering that $\mathbf{A}^{(0)}$ just needs to be in roughly a good region. The symmetry assumption ensures $\mathbf{T}_{mm} = \mathbf{I}_{mm}$ and the the problem reduces to simply the inversion of $\boldsymbol{\Sigma}(t)$. One setting where the move from mean and covariance parameters to linear drift parameters is exact is for univariate problems. For the univariate case, there exist a lot of "of-the-shelf" GP algorithms that can be tested in the free-energy framework.

5.3.4.2 Experiments

In this section, initial values for A and b were constructed from the mean and covariance of some common GP algorithms applied to the double well problem. Figure 5.3 shows the speed of convergence in terms of iterations vs free action. The number of fails gives an indication of the stability of the resulting initialisation. The method ADF2 is the Gaussian projection smoother in section 5.3.2. The method ADF1 is the alternative smoothing approach given in (Särkkä & Sarmavouri, 2011). It is interesting because it is derived directly from the Fokker Planck equations. It is highly unstable, but when it works it shows good results. This suggests a square root form of the algorithm might work well. All Matern kernels had parameters learned using maximum likelihood. The Matern3 algorithm used a Matern type 3 kernel in standard GP regression. It can be seen in figure 5.4 how the Matern3 initialisation struggles to converge and eventually results in covariance instabilities and failure. This effect was even more extreme in the Matern5 kernel and squared exponential kernel (not shown), with failed almost every time. This is a very interesting point because it suggests that the initial A and b rebuilt from these GPs transferred the information in the assumed noise. For all kernels apart from Matern1, the noise assumption was wrong. In a rigorous formulation this would result in a infinite free energy.



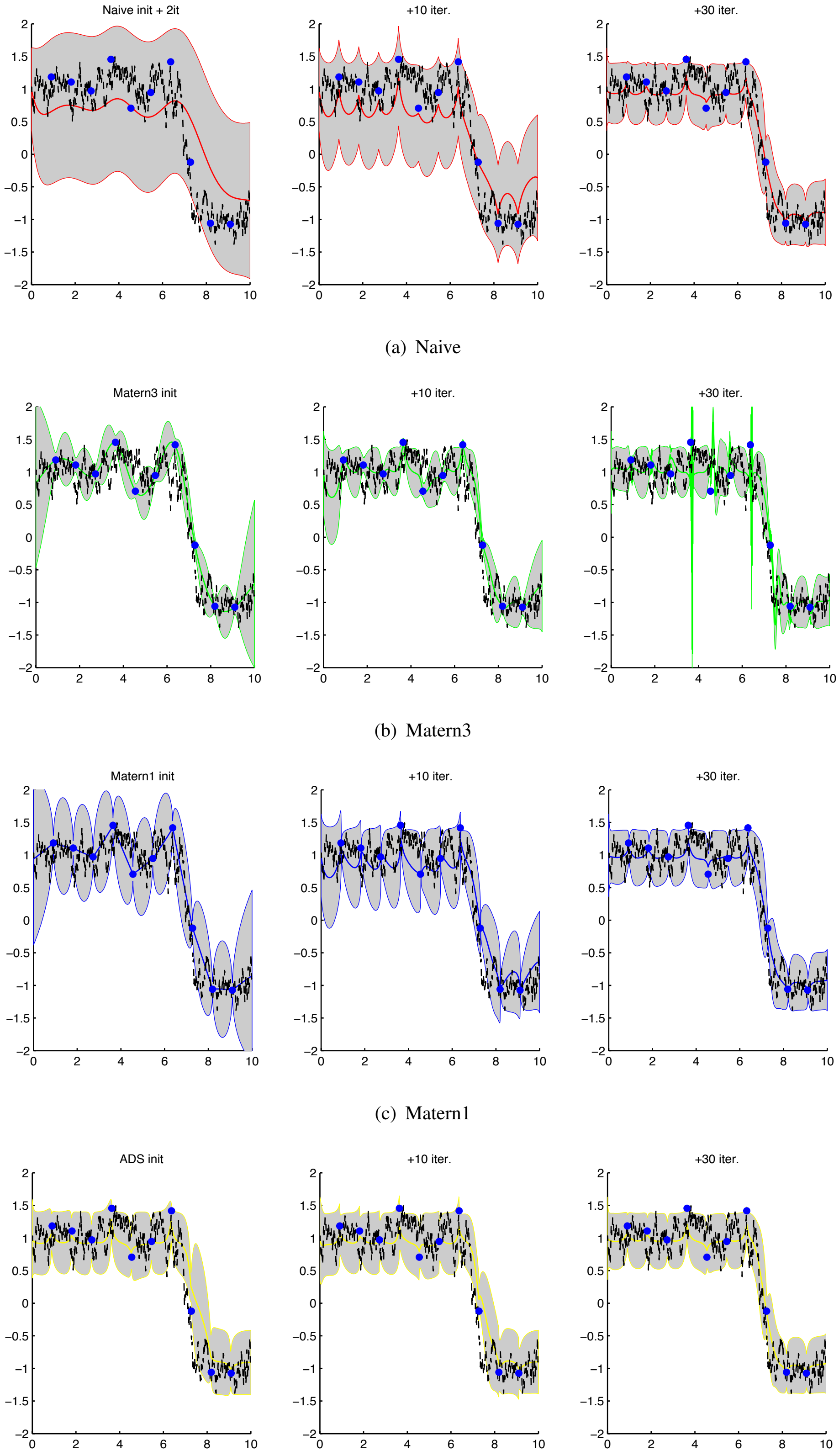
(a) Free-energy vs iterations

(b) Free. vs it. (zoom)

(c) Samp. + VGPS soln.

FIGURE 5.3: Figure (a) shows the path of the free-action of VPG smoothing algorithm initialised with different GP solutions. Results were averaged over 1000 repetitions of sampled data from the double-well system. Figure (b) shows a zoom into the iteration interval $[10, 35]$. Figure (c) shows an example sample path and VGP smoothing solution after convergence.

In experiment it leads to highly unstable initialisations for the variational GP approximation. Matern1 uses the right noise type and converged to the correct solution in most cases. The naive method is the initialisation method used in Vrettas et al. (2010). It is a simple method that chooses a constant value for A, and defines b using a spline fitted to the observations. The key point is its unbiasedness. The Matern3 kernel finds a very precise solution, and it is difficult for the variational algorithm to get out of the probably locally optimal region. By being very unbiased, the naive method apears better at finding globally optimal solutions. Closer inspection of the GPS solution shows how it is close to optimal from the go, in contrast to the naive method. The only region that is finessed is the region around the nonlinear transition. A time plot of the free energy is possible. This suggests an *adaptive* variational GP method that could be combined with GPS. In the adaptive method, the algorithm would focus on regions of high free energy as a function of time. GPS is close to optimal in linear regions around the equilibrium. An adaptive variational GP method could then focus on just the nonlinear region and finesse near optimal solutions in faster time.



(a) Naive

(b) Matern3

(c) Matern1

(d) GPS

FIGURE 5.4: Plots of variational solution after 0, 10 and 30 iterations for different initialisation.

5.4 Implementation

5.4.1 Unscented transform

This section focuses on approximating general transformations of random variables. The underlying principle is that it is better/easier to approximate the distribution of the random variable under a nonlinear transformation than to approximate the transformation itself (Julier & Uhlmann, 2004) . While Monte Carlo methods allow approximations to be generated up to arbitrary accuracy, the difficulty of the sampling process, random sampling errors, and the computational burden of Monte Carlo methods make them often undesirable. Rather than drawing samples randomly from the unmapped distribution, it makes sense to skilfully choose a few samples that encode the core properties of the unmapped distribution. The *Unscented transformation* (UT) (Julier & Uhlmann, 2004) can capture the statistics of the Gaussian exactly. This is done through the use of a set of *sigma points*, whose size scales linearly with the dimension of the random variable. More generally, assume a random variable $\mathbf{x}$ is given, and assume a nonlinear map $\mathbf{h}(\mathbf{x})$ is given such that a random variable $\mathbf{y}$ is defined to be

$$\mathbf{y} = \mathbf{h}(\mathbf{x}). \tag{5.63}$$

The purpose of the unscented transform is to construct a approximation of the distribution of $\mathbf{y}$. A set of sigma points $\mathcal{S} = \{(\mathbf{x}_{(i)}, w_{(i)})\}_{i=0}^{n}$ consists of a set of vectors and their associated weights such that $\sum_i w_{(i)} = 1$, where parenthesis subscripts are used to distinguish the sigma points from discrete-time paths. The sigma points are propagated through the nonlinear map

$$\mathbf{y}_{(i)} = \mathbf{h}(\mathbf{x}_{(i)}), \quad i = 0, 1, \ldots, n. \tag{5.64}$$

Then the first cumulant approximation, $\hat{\boldsymbol{\kappa}}_1$, of $\mathbf{y}$ is a weighted average of the transformed points

$$\hat{\boldsymbol{\kappa}}_1 = \sum_{i=0}^{n} w_{(i)} \mathbf{y}_{(i)}, \tag{5.65}$$

and the second cumulant approximation, $\hat{\boldsymbol{\kappa}}_2$, of $\mathbf{y}$ is a weighted outer product

$$\hat{\boldsymbol{\kappa}}_2 = \sum_{i=0}^{n} w_{(i)} (\mathbf{y}_{(i)} - \hat{\boldsymbol{\kappa}}_1)(\mathbf{y}_{(i)} - \hat{\boldsymbol{\kappa}}_1)^{\mathrm{T}}. \tag{5.66}$$

From $\hat{\boldsymbol{\kappa}}_1$ and $\hat{\boldsymbol{\kappa}}_2$ a Gaussian approximation of $\mathbf{y}$ can be built. It is simple to construct approximations of other functions, such as the diagonal elements of the third cumulant,

$$\hat{\kappa}_3^{(jjj)} = \sum_{i=0}^{n} w_{(i)} (y_{(i)}^{(j)} - \hat{\kappa}_1^{(j)})^3, \quad j = 1, \ldots, d_{\mathbf{y}}. \tag{5.67}$$

There are different ways to choose the optimal sigma points when the input distribution is Gaussian. By moving the skewness in the skew normal density, from the density to the nonlinear transformation, this ensures all the sigma point methods are applicable for approximating skew normal expectations.

5.4.2 Gradient methods

An appealing quality of the variational free-energy approach, in terms of implementation, is that established optimisation algorithms are immediately applicable to the free energy objective. Most nonlinear optimisation problems make use of local quadratic approximations to control step size and direction, improving efficiency over basic gradient-descent type methods. In the recent work of Vrettas (2010), Vrettas et al. (2010), the variational approximation and an efficient suboptimal RBF approximation were implemented using a *scaled conjugate gradient* (SCG) optimisation algorithm, adapted from the *NetLab* version (Nabney, 2002). It is based upon the VGPA algorithm of Archambeau et al. (2007b) and shows improved speed and stability over alternative methods. These alternative methods include the smoothing algorithm in Archambeau et al. (2007a) which uses explicit updates for the variation parameters $\mathbf{A}(t)$ and $\mathbf{b}(t)$, and other *boundary value problem* methods. This section reviews the general SCG algorithm and its motivations. The SCG algorithm and the simpler conjugate gradient (CG) optimisation algorithm are then discussed in the setting of VGPA, and proposals are made for using explicit Hessian conjugates.

5.4.2.1 Conjugated gradients optimisation methods

This section is built out of ideas from (Nabney, 2002, Chapter 2) and Shewchuk (1994). It reviews the basics of gradient based optimisation and the particulars of conjugated and scaled-conjugated gradients. The SCG algorithm shows marked improvements in efficiency over the CG algorithm and *quasi-newton* methods for many *Neural Network* and statistical learning problems (Nabney, 2002). This also witnessed in application to the VGPA algorithm. Here we consider a general objective $f(\boldsymbol{\pi})$ with input vector $\boldsymbol{\pi}$. While the variational parameter $\boldsymbol{\pi}$ notation is used, note that what follows applies to general objective functions and search spaces. All the optimisation methods considered are *iterative* techniques, in that each new point $\boldsymbol{\pi}_{new}$ is built from the recursion

$$\boldsymbol{\pi}_{new} = \boldsymbol{\pi}_i + \alpha_i \boldsymbol{\delta}_i, \tag{5.68}$$

for some step length α_i and step direction $\boldsymbol{\delta}_i$. To ease notation, let $\mathbf{g}_i = \nabla_{\boldsymbol{\pi}_i} f(\boldsymbol{\pi}_i)$ denote the gradient vector of f at step i. Given a step direction $\boldsymbol{\delta}_i$, it makes sense to perform a line-search in the direction $\boldsymbol{\delta}_i$, to find the optimal value for α_i. This is done by solving the fixed point equation (assuming f is locally convex)

$$\frac{\partial f(\boldsymbol{\pi}_{i+1})}{\partial \alpha} = 0. \tag{5.69}$$

Applying the chain rule, we have

$$\frac{\partial f(\boldsymbol{\pi}_{i+1})}{\partial \alpha} = \mathbf{g}_{i+1}^{\mathrm{T}} \frac{\partial \boldsymbol{\pi}_{i+1}}{\partial \alpha} = \mathbf{g}_{i+1}^{\mathrm{T}} \boldsymbol{\delta}_i = 0. \tag{5.70}$$

This implies the new gradient-vector $\mathbf{g}_{i+1}$ is *orthogonal* to the old search direction $\boldsymbol{\delta}_i$. In *gradient descent* or *steepest descent* algorithms the step direction $\boldsymbol{\delta}_i$ is set to be negative gradient $-\mathbf{g}_i$. From equation (5.70), this choice leads to the relation $\boldsymbol{\delta}_{i+1}^{\mathrm{T}} \boldsymbol{\delta}_i = 0$, for consecutive step directions $\boldsymbol{\delta}_{i-1}$ and $\boldsymbol{\delta}_i$. Therefore steepest descent algorithms zig-zag across the search space in consecutively orthogonal directions. While equation (5.69) leads to an optimal choice for α_i, the resulting zig-zag behaviour is generally not the most efficient way to reach a local minimum. A more efficient approach is to choose consecutive step-directions that are orthogonal under a local quadratic approximation of the objective function. This encourages the algorithm to explore the geometry of the objective using the curvature of the local space. Let $\mathbf{H}_i := \nabla_{\boldsymbol{\pi}_i \boldsymbol{\pi}_i^{\mathrm{T}}} f(\boldsymbol{\pi}_i)$ denote the Hessian at step i. Then we are looking for consecutive, *conjugate* directions $\boldsymbol{\delta}_{i-1}$ and $\boldsymbol{\delta}_i$, such that

$$\boldsymbol{\delta}_i^{\mathrm{T}} \mathbf{H}_i \boldsymbol{\delta}_{i-1} = 0. \tag{5.71}$$

The trick in the CG algorithm is to find the vector $\boldsymbol{\delta}_i$ satisfying (5.71) without computing the Hessian $\mathbf{H}_i$. Given current position $\boldsymbol{\pi}$ and search direction δ, let α^* denote the step length minimising $f(\boldsymbol{\pi} + \alpha\boldsymbol{\delta})$. Then, remarkably, the update

$$\boldsymbol{\delta}_{new} = \gamma\boldsymbol{\delta} - \mathbf{g}_{new} \tag{5.72}$$

is conjugate to $\boldsymbol{\delta}$, where

$$\mathbf{g}_{new} = \nabla f(\pi + \alpha^* \boldsymbol{\delta}) \tag{5.73}$$

$$\gamma = \frac{(\mathbf{g}_{new} - \mathbf{g})^{\mathrm{T}} \mathbf{g}_{new}}{\mathbf{g}^{\mathrm{T}} \mathbf{g}}. \tag{5.74}$$

Equation (5.74) is referred to as the *Polak-Ribiere* formula. Equation (5.72) allows new, conjugate, directions to be computed without computing the Hessian $\mathbf{H}_j$. The optimisation method using this type of update is the conjugate gradient (CG) algorithm (Shewchuk, 1994). A highly simplified version of the CG method in shown in algorithm 4. While there are additional termination and quality checks not shown in the code fragment, the core components of the CG algorithm are present. The CG algorithm alternates between a position update involving a line search, and a direction update involving the Polak-Ribiere formula. The main problem with the CG algorithm is the additional computational cost and extra parameter definitions required in the line search. This is avoided in the scaled conjugate method (SCG) of Mller (1993). It can be shown that the optimal step length α^* in line 6 of algorithm 4, is given by the formula

$$\alpha^* = \frac{\mathbf{g}^{\mathrm{T}} \boldsymbol{\delta}}{\boldsymbol{\delta}^{\mathrm{T}} \mathbf{H} \boldsymbol{\delta}}. \tag{5.75}$$

```
**Algorithm:** CG-optimisation (simplified).
% Initialise variables;
$\boldsymbol{\pi} = \boldsymbol{\pi}_0$; $\mathbf{g} = \nabla f(\boldsymbol{\pi})$; $\boldsymbol{\delta} = -\mathbf{g}$;
repeat
    % Perform line search to update position;
    $\alpha^* = \min_\alpha f(\boldsymbol{\pi} + \alpha\boldsymbol{\delta})$;
    $\boldsymbol{\pi}_{new} = \boldsymbol{\pi} + \alpha^*\boldsymbol{\delta}$;
    % Use Polak-Ribiere formula to update search direction;
    $\mathbf{g}_{new} = \nabla f(\boldsymbol{\pi}_{new})$;
    $\gamma = (\mathbf{g}_{new} - \mathbf{g})^{\mathrm{T}}\mathbf{g}_{new}/\mathbf{g}^{\mathrm{T}}\mathbf{g}$;
    $\boldsymbol{\delta}_{new} = \gamma\boldsymbol{\delta} - \mathbf{g}_{new}$;
until covergence, i.e. $\mathbf{g}^{\mathrm{T}}\mathbf{g} = 0$;
```

Algorithm 4: Simplified conjugate gradient algorithm.

The reason this is not used in the CG-algorithm is because it is assumed that the Hessian $\mathbf{H}$ is either unavailable or costly to compute. In the SCG algorithm, the denominator in equation (5.75) is approximated by

$$\boldsymbol{\delta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\delta} \approx \boldsymbol{\delta}^{\mathrm{T}}\Big(\nabla f(\boldsymbol{\pi} + \sigma_0\frac{\boldsymbol{\delta}}{||\boldsymbol{\delta}||_2}) - \nabla f(\boldsymbol{\pi})\Big)\frac{||\boldsymbol{\delta}||_2}{\sigma_0}, \tag{5.76}$$

for some small positive quantity σ_0. To ensure the denominator in equation (5.75) is positive definite, a *ridge* $\beta||\boldsymbol{\delta}||^2$ is added at each step, where the value β changes with each iteration. There is an additional *comparison ratio* check (Nabney, 2002) to ensure the approximation is close to quadratic, and also all the rudimentary checks of the CG algorithm. Ignoring these, a simplified SCG method is shown in algorithm 5. The core components is the approximation of the quadratic form $\boldsymbol{\delta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\delta}$, and the Polak-Ribiere update (as in CG). If a minimum has not

```
**Algorithm:** SCG-optimisation (simplified).
% Initialise variables;
$\boldsymbol{\pi} = \boldsymbol{\pi}_0$; $\mathbf{g} = \nabla f(\boldsymbol{\pi})$; $\boldsymbol{\delta} = -\mathbf{g}$;
repeat
    % Compute (approximate) optimal step length to update position;
    $\Delta \approx \boldsymbol{\delta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\delta}$;
    $\alpha^* = \mathbf{g}^{\mathrm{T}}\boldsymbol{\delta}/\Delta$;
    $\boldsymbol{\pi}_{new} = \boldsymbol{\pi} + \alpha^*\boldsymbol{\delta}$;
    % Use Polak-Ribiere formula to update search direction;
    $\mathbf{g}_{new} = \nabla f(\boldsymbol{\pi}_{new})$;
    $\gamma = (\mathbf{g}_{new} - \mathbf{g})^{\mathrm{T}}\mathbf{g}_{new}/\mathbf{g}^{\mathrm{T}}\mathbf{g}$;
    $\boldsymbol{\delta}_{new} = \gamma\boldsymbol{\delta} - \mathbf{g}_{new}$;
until covergence, i.e. $\mathbf{g}^{\mathrm{T}}\mathbf{g} = 0$;
```

Algorithm 5: Simplified scaled-conjugate gradient algorithm.

been reached in $d_{\boldsymbol{\pi}}$ iterations, where $d_{\boldsymbol{\pi}}$ denotes the dimension of the space in which $\boldsymbol{\pi}$ lies, then the algorithm restarts at the current location $\boldsymbol{\pi}$ with step direction $\boldsymbol{\delta} = -\nabla f(\boldsymbol{\pi})$. Note that alternative search-direction updates can be used in both algorithms and that they both avoid the use of the Hessian.

5.4.2.2 Free-energy Hessians for VGPA

The gradients of the Lagrangian $\hat{\mathcal{L}}$ used in Archambeau et al. (2007b) are given in equations (3.28) and (3.29). When the algorithm is implemented, the variables are defined on a discrete time grid $0 = t_0 < \cdots < t_N = T$. The variational variables $\mathbf{A}(t)$ and $\mathbf{b}(t)$ can therefore be considered as discrete time functions such that $\mathbf{A} : [0 : N] \to \mathbb{R}^{d_\mathbf{x} \times d_\mathbf{x}}$ and $\mathbf{b} : [0 : N] \to \mathbb{R}^{d_\mathbf{x}}$. The gradients in equations (3.28) and (3.29) can thus be written

$$\begin{aligned} \nabla_{\mathbf{A}_i} \hat{\mathcal{L}} &= \nabla_{\mathbf{A}_i} E_{sde}(i) + \boldsymbol{\Sigma}(i)\boldsymbol{\Lambda}^{\mathrm{T}}(i) + \boldsymbol{\Lambda}(i)\boldsymbol{\Sigma}(i) + \boldsymbol{\nu}(i)\boldsymbol{\mu}^{\mathrm{T}}(i) && (5.77)\\ \nabla_{\mathbf{b}_i} \hat{\mathcal{L}} &= \nabla_{\mathbf{b}_i} E_{sde}(i) + \boldsymbol{\nu}(i). && (5.78) \end{aligned}$$

The Hessian $\mathbf{H}$ of the Lagrangian $\hat{\mathcal{L}}$ is a huge $d_\mathbf{x}N(1+d_\mathbf{x}) \times d_\mathbf{x}N(1+d_\mathbf{x})$ dimensional matrix. But note the explicit forms for the E_{sde} gradients in (5.77) and (5.78), for state-independent diffusion $\mathbf{D}(\mathbf{x}, t) = \mathbf{D}_t$, are given by

$$\begin{aligned} \nabla_{\mathbf{A}_i} E_{sde}(i) &= \mathbf{D}_i^{-1}\Big(\mathbf{A}(i) - \big\langle \nabla_\mathbf{x} \mathbf{f}^{\mathrm{T}}(\mathbf{x}, i)\big\rangle\Big)\boldsymbol{\Sigma}(i) + \nabla_{\mathbf{b}_i} E_{sde}(i)\boldsymbol{\mu}^{\mathrm{T}}(i) && (5.79)\\ \nabla_{\mathbf{b}_i} E_{sde}(i) &= \mathbf{D}_i^{-1}\Big(\mathbf{A}(i)\boldsymbol{\mu}(i) + \mathbf{b}(i) - \big\langle \mathbf{f}(\mathbf{x}, i)\big\rangle\Big). && (5.80) \end{aligned}$$

Taking second-order derivatives of $\hat{\mathcal{L}}$, for all $i \neq j$ we obtain

$$\begin{aligned} \nabla_{\mathbf{A}_i}\nabla_{\mathbf{A}_j} \hat{\mathcal{L}} &= \mathbf{0}_{d_\mathbf{x}^2 \times d_\mathbf{x}^2} && (5.81)\\ \nabla_{\mathbf{A}_i}\nabla_{\mathbf{b}_j} \hat{\mathcal{L}} &= \mathbf{0}_{d_\mathbf{x}^2 \times d_\mathbf{x}} && (5.82)\\ \nabla_{\mathbf{b}_i^{\mathrm{T}}}\nabla_{\mathbf{b}_j} \hat{\mathcal{L}} &= \mathbf{0}_{d_\mathbf{x} \times d_\mathbf{x}}. && (5.83) \end{aligned}$$

Therefore the majority of entries in the Hessian are in fact zero. The Hessian $\mathbf{H}$ can, thus, be written as a block matrix composed of four diagonal blocks

$$\mathbf{H} = \left(\begin{array}{c|c} \mathbf{H}_{\mathbf{AA}} & \mathbf{H}_{\mathbf{Ab}} \\ \hline \mathbf{H}_{\mathbf{bA}} & \mathbf{H}_{\mathbf{bb}} \end{array}\right), \qquad (5.84)$$

where

$$\begin{aligned} \mathbf{H}_{\mathbf{AA}} &:= \mathrm{diag}\Big((\nabla_{\mathbf{A}_i}\nabla_{\mathbf{A}_i})_{i=0}^N\Big)\hat{\mathcal{L}} && (5.85)\\ \mathbf{H}_{\mathbf{Ab}} &:= \mathrm{diag}\Big((\nabla_{\mathbf{A}_i}\nabla_{\mathbf{b}_i})_{i=0}^N\Big)\hat{\mathcal{L}} && (5.86)\\ \mathbf{H}_{\mathbf{bb}} &:= \mathrm{diag}\Big((\nabla_{\mathbf{b}_i}\nabla_{\mathbf{b}_i})_{i=0}^N\Big)\hat{\mathcal{L}} && (5.87) \end{aligned}$$

with diagonal elements

$$\nabla_{\mathbf{A}_i}\nabla_{\mathbf{A}_i}\hat{\mathcal{L}} = \mathbf{D}_i^{-1} \otimes \left(\mathbf{\Sigma}(i) + \boldsymbol{\mu}(i)\boldsymbol{\mu}^{\mathrm{T}}(i)\right) \tag{5.88}$$

$$\nabla_{\mathbf{A}_i}\nabla_{\mathbf{b}_i}\hat{\mathcal{L}} = \mathbf{D}_i^{-1} \otimes \boldsymbol{\mu}(i) \tag{5.89}$$

$$\nabla_{\mathbf{b}_i}\nabla_{\mathbf{b}_i}\hat{\mathcal{L}} = \mathbf{D}_i^{-1}. \tag{5.90}$$

We know the Hessian of $\hat{\mathcal{L}}$ is positive definite because the free-energy is quadratic in $\mathbf{A}(i)$ and $\mathbf{b}(i)$, and equations (5.88), (5.89), and (5.90) confirm this. Let $\boldsymbol{\delta}_{\mathbf{A}(i)}$ and $\boldsymbol{\delta}_{\mathbf{b}(i)}$ denote the components of $\boldsymbol{\delta}$ corresponding to $\mathbf{A}(i)$ and $\mathbf{b}(i)$ in the scaled conjugate gradient algorithm, such that $\boldsymbol{\delta}_{\mathbf{A}(i)}$ is a matrix and $\boldsymbol{\delta}_{\mathbf{b}(i)}$ a vector. Then we have

$$\boldsymbol{\delta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\delta} = \sum_{i=0}^{N} \mathrm{tr}\left\{\boldsymbol{\delta}_{\mathbf{A}(i)}^{\mathrm{T}}\left(\mathbf{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}\right)\boldsymbol{\delta}_{\mathbf{A}(i)}\mathbf{D}_i^{-1}\right\} \tag{5.91}$$

$$+2\mathrm{tr}\left\{\boldsymbol{\delta}_{\mathbf{A}(i)}^{\mathrm{T}}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\delta}_{\mathbf{b}(i)}^{\mathrm{T}}\mathbf{D}_i^{-1}\right\} + \boldsymbol{\delta}_{\mathbf{b}(i)}^{\mathrm{T}}\mathbf{D}_i^{-1}\boldsymbol{\delta}_{\mathbf{b}(i)}. \tag{5.92}$$

Thus the quadratic form involving the Hessian can be computed from objects already being used in the VGPA algorithm. This suggests replacing the approximation in SCG with its true value. The idea would also applicable to *Newton's method*, replacing the need for a quasi-Newton alternative. It might also be possible to introduce locally skew-normal approximations instead of simply quadratic ones and to extend the SCG algorithm using the ideas of Minka (2000). Future investigations will examine how well these all work in practice.

5.5 Discussion

This chapter has attempted to place the variational GP method in the context of other GP learning methods and to consider alternative GP methods from a free energy perspective. An exact Hessian update has also been proposed for gradient methods.

The connections between the variational GP algorithm and other algorithms were drawn for several reasons. Firstly, general GP and kernel methods can easily capture higher order noise correlations. For state space model methods to do this requires an increase in the dimension of the problem. For example, to obtain the smoothing properties of the squared exponential would require a state space model of infinite dimension. It would be beneficial for applications if there was a way of adapting the state space model to integrate higher order noise correlations without increasing dimension, maybe through the integration of more general linear GP methods. A related investigation would work on the connection between general covariance functions and linear time invariant state space models. While it is well known how regularisation operators can be used to build covariance operators and how these lead to state space models and GPs, respectively, the full algorithmic details of the connection have not been formalised. State space model methods can be very fast for uncorrelated noise, taking full advantage of the independence of observations given the hidden states. Given the growth in dimension to capture more general noise correlations, there must be a point at which the use of covariance functions and the kernel matrix inversion becomes more beneficial than the use of the Kalman smoother. Another reason for identifying the connections is the fact that kernel and GP methods can only be built from linear differential operators. The variational GP algorithm presents a possible way to incorporate nonlinear information into GP and kernel methods. It is not known yet how beneficial this connection would be, or whether or not the whole problem could be dealt with in the Variational GP framework. To strengthen the connection further would require the extension of the variational algorithm to spatio and spatio-temporal methods. While this would require a discretisation of a high dimensional continuous input, it is possible that empirical input spaces such as nearest neighbour graphs and differential operator approximations such as the graph Laplacian could be used to avoid such computationally expensive grid like discretisations.

Chapter 6

Discussion and Outlook

6.1 Machine learning in continuous time

To perform temporal state estimation on a digital computer requires the discretisation of a finite-length interval of time. While the graining of this discretisation can be made arbitrarily small, this has significantly differing consequences for the performance of each algorithm. The temporal state estimation problem has a unique structure, and applying established algorithms to a discretisation of the model it not always effective. Ideally, each class of machine learning algorithm will have an interpretation tailored to the requirements of state estimation in partially observed diffusion processes. Traditional machine learning algorithms are designed for finite dimensional states. When applied to state-space models restricted to a discretisation of time, these finite dimensional algorithms require a closed form for the transition probabilities of the discretised model. For general nonlinear state space models this is not possible, and the transition probabilities can only be formalised as the solutions to a set of generally intractable stochastic partial differential equations. Therefore the common approach to adapting a machine learning algorithm to a continuous-discrete state estimation problem is to use linear approximations of the transition probabilities. This equates to a simplification of the prior, wasting important information in the prior, as well as leading to numerical instabilities in implementation. *Continuous-time algorithms in machine learning are algorithms that are not restricted to first order linear approximations*. These algorithms are realised in one of two ways, depending on whether the approximate inference method is stochastic or deterministic. Stochastic approximate inference methods utilise higher order approximations of transition probabilities implicitly through the use of higher order approximations of the underlying process. This enables more prior information to be incorporated into the algorithm without increasing the granularity of the discretisation. In contrast, deterministic approximate inference methods restrict themselves to well behaved expectations, as opposed to non differentiable sample paths. Finite dimensional deterministic methods can be applied to first order linear approximations of the model. The well behaved nature of expectations allows the maximum step size of the time discretisation to be taken to zero, reducing the error of the first order approximation to a negligible level, and resulting in continuous-time formulations of the original finite dimensional algorithms. Continuous time parameters characterising the deterministic approximate inference solution are presented in the form of differential equations that are substantially easier to solve than the original stochastic partial differential equation. An infinite dimensional formulation of an algorithm, if it exists, characterises how the algorithm will perform as the step size tends to zero. While this requires the delicate handling of probability measures on infinite dimensional spaces, it enables us to choose parameters for the algorithm that are independent of dimension.

6.1.1 Path integral formulation

Rather than the linear discretisation used in section 2.5.4, the continuous-discrete posterior P_{post} can be written in continuous *path integral* form

$$\frac{dP_{post}}{dW}(\mathbf{x}) \propto \exp\left(- \int_0^T \left\{ \frac{1}{2}||\mathbf{f}(\mathbf{x},t)||^2_{\mathbf{D}(\mathbf{x},t)} dt - \mathbf{f}^{\mathrm{T}}(\mathbf{x},t)\mathbf{D}(\mathbf{x},t)d\mathbf{x} \right\} - U(\mathbf{x},t) \right) \tag{6.1}$$

where the second integral is an Itô stochastic integral and $U(\mathbf{x},t)$ is given by equation (3.12). Any algorithm designed to approximate P_{post} can only do this on a discretised time grid of size N. How the algorithm performs as a function of N is paramount to its consideration as a continuous-time algorithm. A reason for poor performance in the limit $N = \infty$ can be down to the fact that an algorithm does not have a natural formulation in infinite dimensions. In Beskos et al. (2011) the Hybrid Monte-Carlo algorithm is extended to Hilbert spaces. This allows us to work with the posterior given in (6.1), at least up to the stage of implementation. The core components of the HMC algorithm, namely, the *Hamiltonian flow*, the *numerical integrator*, and the *accept/reject rule* all have extensions to the continuous time setting. The benefit of all this, is that it allows important parameters of the algorithm to be calibrated independently of N. Experimental work in Beskos et al. (2011) has shown how the performance of the finite dimensional HMC degrades as $N \to \infty$, but the performance of the infinite dimensional version is indifferent to N. The performance of the algorithm is measured using the acceptance probability of new samples, being an indication of how well the algorithm is exploring the potential landscape. The infinite dimensional algorithm obviously increases in computational complexity when implemented on a time-discretisation of increasing size, but its consistent acceptance rate allows it to more quickly acquire a predefined number of accepted samples. Having a formulation that is well-defined in the continuous-time limit and the fact that the finite-dimensional HMC algorithm can be implemented using higher-order SDE integrators (Restrepo, 2008), with the complexity of the HMC algorithm commensurating with the complexity of the integrator, lets us conclude that HMC, with the correct tuning, can be considered a continuous-time algorithm.

6.1.2 Free action methods

Methods in this section receive special attention because they are the approximate inference methods most closely related to the variational inference scheme considered here.

6.1.2.1 Variational filtering

Variational filtering (Friston, 2008b) is a free action approach that avoids using a fixed form for the variational density. It assumes the sensor data is a continuous time *observation process* $\{\mathbf{y}_t\}_{t\in[0,T]}$ with observation model $p(\mathbf{y}_t|\mathbf{x}_t)$. The focus is on estimation of the conditioned marginal density of $\mathbf{x}_t$ on $\mathbf{y}_t$. The free action is expressed in integral form

$$\mathcal{A}_{vf}(q) = \int_0^T dt \big\langle V(\mathbf{x},t) + \log q(\mathbf{x},t) \big\rangle_{q(\mathbf{x},t)} \tag{6.2}$$

where $V(\mathbf{x},t) = -\log p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t)$. This immediately presents a problem with interpreting variational filtering in the context studied here. The marginal density $p(\mathbf{x}_t)$ of the prior at time t is needed in explicit form. We know this requires solution of the Fokker-Planck equation, and so it is difficult to see how variational filtering can pose a computationally efficient approximate inference method. Continuing with the formulation as in Friston (2008b), the free energy $\mathcal{F}_{vf}(q)$ can be minimised with respect to q under the constraint $\int q(\mathbf{x},t)d\mathbf{x} = 1$, to give

$$q(\mathbf{x},t) = \frac{1}{Z}\exp\big(-V(\mathbf{x},t)\big), \tag{6.3}$$

where Z ensures $\int q(\mathbf{x},t)d\mathbf{x} = 1$. Inserting $V(\mathbf{x},t)$ into equation (6.3) takes us directly back to

$$q(\mathbf{x},t) = \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{y}_t)}. \tag{6.4}$$

So the marginal density of the variational approximation in variational filtering is only conditioned on the data at that particular instant of time. In Friston (2008b) *generalised coordinates* are used to enforce dependencies between time separated variables and to support conditional densities on paths. The most intuitive way to think of generalised coordinates is to assume the state $x(t)$ is deterministic and smooth. Then the differentials $x', x'', x''', \ldots$ fully characterise the path $x(t)$. This obviously does not fit the approach used here. It is conjectured here that these properties ensure that $q(\mathbf{x},t)$ is, indeed, also conditioned on non adjacent data that is reasonably close in time. Under the assumption of a continuous stream of sensor input, this should ensure that the solution is an approximation of the smoothing solution in a local sense. The use of generalised coordinates is another reason why it is difficult to interpret variational filtering in the context studied here. In the *ensemble* approach used in Friston (2008b) the solution $q(\mathbf{x},t)$ is considered an ensemble density that flows on the variational manifold defined by $V(\mathbf{x},t)$. While the manifold is changing with time and the ensemble cannot settle to a steady state, a frame of reference that moves with the manifolds topology is constructed such that the ensemble is in equilibrium. The same idea is used in section (4.2) for constructing well behaved time-varying variational drifts. It is not possible for the scheme in Friston (2008b) to be represented in a fully rigorous fashion. Instead, consider the following equation

$$d\mathbf{x}_t = \boldsymbol{\mu}_t + V_{\mathbf{x}}(\mathbf{x}, t) + d\mathbf{w}_t, \quad \text{s.t.} \quad V_{\mathbf{x}}(\boldsymbol{\mu}_t, t) = 0, \tag{6.5}$$

where $V_{\mathbf{x}}(\mathbf{x}, t) := \nabla_{\mathbf{x}} V(\mathbf{x}, t)$. Using this to define the motion of the ensemble, equation (6.5) is integrated forward in time multiple times and the resulting sample distribution is used to approximate $q(\mathbf{x}, t)$. The difference between equation (6.5) and equation (16b) of Friston (2008b) is the use of generalised coordinates. While equation (6.5) looks similar to a *Langevin algorithm* (Stramer & Tweedie, 1999), there are some distinct differences. Firstly, the potential $V(\mathbf{x}, t)$ is time varying, therefore the corresponding densities are non stationary. This is a key difference between using ensemble methods for static models and using them for temporal models. The second difference between equation (6.5) and the *Langevin algorithm*, is the use of the mode $\boldsymbol{\mu}_t$. The free-energy manifold is constantly changing with time and the use of $\boldsymbol{\mu}_t$ is proposed in Friston (2008b) to ensure the ensemble of particles clouds around the mode. To implement the variational filter, time is discretised and the stochastic integration method of Ozaki (1993) is used, but any stochastic integration scheme, of any order, could be used. Variational filtering shows equivalent accuracy compared to DEM with Laplace approximation (see following section) on a linear model, but is an order of magnitude slower. It shows competitive results with particle filter. The key difference between variational filtering and the variational smoothing framework formulated here is how in variational filtering the variational approximation is not given a fixed form. The principle advantage of variational filtering over conventional methods is that it is easily extended to more complex models than shallow state space models studied here.

6.1.2.2 DEM with Laplace approximation

The variational filtering method of the previous section can approximate conditionals on states with any form. In contrast, *DEM with Laplace approximation* (Friston et al., 2008) assumes a fixed Gaussian form for the ensemble $q(\mathbf{x}, t)$. The *dynamic* (D) part of DEM is, in fact, all we are concerned with here, being the step that updates the approximation of the conditional density on states. Assume the ensemble $q(\mathbf{x}, t)$ is a Gaussian with mean $\boldsymbol{\mu}_t$ and covariance $\boldsymbol{\Sigma}_t$. Ignoring constants, the Laplace approximation of the free-action is given by

$$\mathcal{A}_{lap}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) = \int_0^T V(\boldsymbol{\mu}_t, t) + \frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}_t V_{\mathbf{xx}}(\boldsymbol{\mu}_t, t)\} - \frac{1}{2}\log|\boldsymbol{\Sigma}_t| dt \tag{6.6}$$

where $V_{\mathbf{xx}}(\mathbf{x}, t) = \nabla_{\mathbf{x}}^{\mathrm{T}} \nabla_{\mathbf{x}} V(\mathbf{x}, t)$. Though in (Friston et al., 2008), the quantity $\frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}_t V_{\mathbf{xx}}(\boldsymbol{\mu}_t, t)\}$ only appears in the *internal energy* of the unknown parameters, through mean-field coupling and does not appear in the free-action of the conditional. The covariance is set to the inverse Hessian of the free energy evaluated at $\boldsymbol{\mu}_t$, i.e. $\boldsymbol{\Sigma}_t = V_{\mathbf{xx}}^{-1}(\boldsymbol{\mu}_t, t)$. The mean $\boldsymbol{\mu}_t$ is then learned by integrating the equation

$$\dot{\boldsymbol{\mu}}_t = \kappa V_{\mathbf{x}}(\boldsymbol{\mu}_t, t) \tag{6.7}$$

forward in time, for some dampening constant $\kappa > 0$. This is done numerically using the approach of Ozaki (1992). The main differences between the exposition given here and the one in Friston et al. (2008) are: Firstly, equation (6.7) is done in generalised coordinates and involves higher orders of motion of the mode $\boldsymbol{\mu}', \boldsymbol{\mu}'', \boldsymbol{\mu}''', \ldots$. It is still unclear if this is can be translated into our framework, but it should involve introducing higher order derivatives of the state variables and coupling them in the right way. Secondly, in Friston et al. (2008) equation (6.6) is coupled with em internal free-actions involving unknown parameters. Importantly, states (D), parameters (E), and hyper parameters (M), are learned iteratively in the DEM algorithm. By reducing the algorithm to just the D step, as above, we are restricted to only one pass for the mode which is essentially just a filter. In DEM, the additional uncertainty in the parameters and hyper parameters and the additional D steps that follow from DEM give the algorithm the chance to make multiple forward integrations of equation (6.7). This, the uncertainty in the parameters, and the additional generalised coordinates, should result in the end solution $\boldsymbol{\mu}_t$ being more of a *smoothing* mode. In experiment, DEM is shown to outperform the extended Kalman filter and have comparative performance with the particle filter on conditional density estimation in a nonlinear convolution model (Friston et al., 2008). The argument in Friston et al. (2008) is that the Kalman filter does not have access to generalised coordinates of motion like DEM, and does not have a free-form density like particle filtering. This suggests that the generalised coordinates remedy the fixed form nature of the Laplace approximation somewhat.

In summary of DEM and variational filtering, both methods propose interesting alternatives to the free-action method studied in this thesis. They are application driven, with origins rooted in neuroscience. Their neuroscience connections leads to interesting discussions about the use of free-action in the brain (Friston et al., 2006). But their formulations often skip over some very important issues from a mathematical perspective. It is felt they are needed to give a full picture of current continuous-time free-action methods, but it is a struggle to integrate their formulation into the one taken here. Future investigations will look at taking them back to first principles and rebuilding the methods from an Itô calculus grounded machine learning perspective.

6.1.2.3 Generalised filtering

An extension of variational filtering, *generalised filtering* (Friston et al., 2010) incorporates posterior approximations of unknown parameters into the conditional density through *slow moving* auxiliary dynamics. The idea of incorporating unknown parameters as slow moving states can be done seamlessly in a hierarchal dynamic model. The are two problems with the approach to generalised filtering taken in Friston et al. (2010), with regards to interpreting in the paradigm taken here. The first is the use of generalised coordinates, as is discussed in the two previous section. The second is the form they choose for their approximation of the free action. Assume the state variable $\mathbf{x}_t$ is made of authentic states and unknown time varying parameters. As in DEM (Friston et al., 2008), the Laplace assumption is made, such that the free energy (time differential of free action), is approximated by by

$$\hat{\mathcal{F}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, t) = V(\boldsymbol{\mu}_t, t) + \frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}_t V_{\mathbf{xx}}(\boldsymbol{\mu}_t, t)\} - \frac{1}{2}\log|\boldsymbol{\Sigma}_t|. \tag{6.8}$$

Optimising with respect to $\boldsymbol{\Sigma}_t$ yields, just as in the standard Laplace approximation, $\boldsymbol{\Sigma}_t = V_{\mathbf{xx}}^{-1}(\boldsymbol{\mu}_t, t)$. Inserting this back into equation (6.8), leads to a free energy approximation that is a function of (and only of) the conditional mean

$$\hat{\mathcal{F}}(\boldsymbol{\mu}_t, t) = V(\boldsymbol{\mu}_t, t) + \frac{1}{2}\log|V_{\mathbf{xx}}(\boldsymbol{\mu}_t, t)|. \tag{6.9}$$

The idea is then to minimise $\hat{\mathcal{F}}(\boldsymbol{\mu}_t, t)$ over $\boldsymbol{\mu}_t$. Generalised filtering showed marked improvements over DEM on some particular models for detecting visual motion-dependent responses in the brain (Friston et al., 2010), and more generally showed performance comparative to DEM. While it is not known if this reasonable performance is down to the use of generalised coordinates or the use of regularisation enforcing priors, it is conjectured here that there exists a fundamental flaw with the use of equation (6.9) as an objective function. Without any additional priors, the $\boldsymbol{\mu}_t$ minimising $\hat{\mathcal{F}}(\boldsymbol{\mu}_t, t)$ will always move to the inflection points of $V(\mathbf{x}, t)$. Recalling that the inflection points of a function are point where the curvature is zero, we see that the inclusion of $\log|V_{\mathbf{xx}}(\boldsymbol{\mu}_t, t)|$ ensures that $\boldsymbol{\mu}_t$ can achieve infinitely negative values of $\hat{\mathcal{F}}(\boldsymbol{\mu}_t, t)$ simply by moving towards inflection points. While this is bad on its own, it is made worse by the fact that inflection points are generally the furthest points on a curve from the critical points which general have high levels of curvature. This is simply seen through a static example. Let $V(x)$ be the energy function that follows from the assimilation of one observation with a linear observation model into the double well steady state, given by

$$V(x) = \frac{(x-y)^2}{2} + bx^2(x^2 - 2\theta). \tag{6.10}$$

Differentiating twice yields

$$V_{xx}(x) = 1 + 4b(3\mu^2 - \theta). \tag{6.11}$$

Therefore the free energy approximation in equation (6.9) for this static example is given by

$$\hat{\mathcal{F}}(\mu) = \frac{(\mu-y)^2}{2} + b\mu^2(\mu^2 - 2\theta) + \frac{1}{2}\log|1 + 4b(3\mu^2 - \theta)|. \tag{6.12}$$

In figure 6.1 the free energy approximation $\hat{\mathcal{F}}(\mu)$ (red) is compared to the negative log likelihood $V(\mu)$ (blue). In a standard *post hoc* Laplace approximation (Tierney & Kadane, 1986), $V(\mu)$ would be minimised over μ and a Gaussian would be built at the global minimum. In contrast,

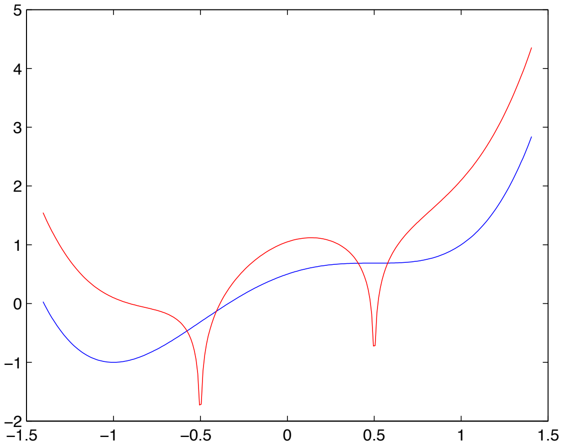


FIGURE 6.1: Plot of $\hat{\mathcal{F}}(\mu)$ (red) and $V(\mu)$ (blue) for double well steady state prior with single linear observation $y = -1$, $\theta = 1$ and $b = 1$.

in the variational Laplace approach of Friston et al. (2010) the objective function $\hat{\mathcal{F}}(\mu)$ has infinitely negative peaks at the inflection points of $V(\mu)$. While is not known how this effect is avoided in generalised filtering, it is conjectured here that the objective function in equation (6.9) is fundamentally flawed.

While the above expositions of variational filtering, DEM with Laplace assumption, and generalised filtering have been quite critical, the methods have some very useful components that should be considered in more detail. Firstly, the use of higher motions to capture time correlations between non adjacent variables. In many real world dynamic systems, such as the ones studied in neuroscience, the noise element in many cases results from interference with other closely connected dynamical systems not included in the model. These noise processes can certainly not always be considered to have independent increments, as with the Wiener process. Secondly, the DEM structure of having a dynamic step for learning the conditional density over states, an expectation step for learning static parameters, and a maximisation step for learning hyper parameters is a natural extension of the standard EM algorithm (Ghahramani & Roweis, 1999) to the *triple estimation* problem. There is no need to assume a Laplace, or even Gaussian approximation in DEM, and the variational GP approximation and the higher order approximations proposed in this thesis fit naturally into the DEM framework and are likely to provide improved results over methods based on the Laplace assumption. Thirdly, the idea of integrating unknown parameters as slowly moving states is an important point. This idea has been implemented in a recent approach (Havlicek et al., 2011) used to model neuronal responses in fMRI. It is a recursive assumed density smoothing approach which has been shown to outperform DEM on a Lorenz attractor model with unknown parameters and initial states. The multiple forward and backward iterations used by the variational GP smoothing algorithm of Archambeau et al. (2007a) should show improvements over the one forward and backward pass approach of assumed density smoothing. Parameters required to be non zero can still be learned with a Gaussian approximation through the use of log transformations.

6.1.3 Towards continuous-time EP

The one deterministic algorithm that doesn't appear to have a natural formulation in continuous-time, as yet, is Expectation Propagation. The variational framework of chapter (3), the projection filter of chapter (3.4), and in some ways the assumed density smoother, all integrate the Fokker-Planck equation into the workings of the algorithm. This equation is at the core of the exact solution. Given a grid of times intersecting the observation times, a good EP-smoothing algorithm would first compute $p(\mathbf{x}_i|\mathbf{x}_{i-1})$ for every i using the Kolmogorov equation (2.25) and then apply the generic EP-smoothing algorithm 2.5.3. This would convert the continuous-time algorithm to a form that EP can accept. But, or course, $p(\mathbf{x}_i|\mathbf{x}_{i-1})$ can generally not be computed and stored for later use, which is why the variational and projection filter algorithms have to use the Kolmogorov equations "on-the-fly". The EP-smoothing algorithm of section 2.5.3 can be given a definite temporal scheduling structure; updating forward messages while keeping backward messages fixed, and vice-versa. We know that this symmetric message passing formulation holds in continuous-time, and the variational algorithm utilises this with its path/control structure of a forward pass to update moments and a backward pass to update Lagrange multipliers. It feels like it should be possible, and time has been spent trying, to obtain a continuous-time limit formulation in the same way as was done for ADF, ADS and the projection smoother, especially given how ADF seeded the EP algorithm. The projection smoother provides evolution equations in moment and canonical form, something a continuous-time EP algorithm is likely to require given how the discrete-time version can be formulated fully in terms of canonical parameters. The exact continuous-time smoothing posterior $p_s(\mathbf{x},t) \propto p_F(\mathbf{x},t)\psi(\mathbf{x},t)$ is proportional to the product of the filter $p_F(\mathbf{x},t)$ and the likelihood $\psi(\mathbf{x},t)$, as in the the discrete-time setting. Therefore it is natural to assume that a continuous-time EP smoothing algorithm will derive an approximation $q(\mathbf{x},t) \propto q_{\boldsymbol{\alpha}}(\mathbf{x},t)q_{\boldsymbol{\beta}}(\mathbf{x},t)$, where $q_{\boldsymbol{\alpha}}(\mathbf{x},t)$ and $q_{\boldsymbol{\beta}}(\mathbf{x},t)$ represent the information coming from the left and right, respectively. In exponential form, the messages $q_{\boldsymbol{\alpha}}(\mathbf{x},t)$ and $q_{\boldsymbol{\beta}}(\mathbf{x},t)$ can be written

$$q_{\boldsymbol{\alpha}}(\mathbf{x},t) = \exp\Big(\boldsymbol{\alpha}_t^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\Big) \tag{6.13}$$

$$q_{\boldsymbol{\beta}}(\mathbf{x},t) = \exp\Big(\boldsymbol{\beta}_t^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\Big), \tag{6.14}$$

where $\boldsymbol{\phi}(\mathbf{x})$ is a common sufficient statistic and $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are canonical parameters that evolve over (continuous) time. The core characteristic of the EP-smoothing algorithm is its symmetrical form, therefore we want to try and obtain a continuous-time algorithm that emulates this. The idea of a "posterior" drift is apparent in all the algorithms and the link with optimal control suggests that an approximate drift will accompany the algorithm, and will be of the form

$\mathbf{g}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, t)$ using the forward and backward messages as controls. It is not known if this will be an additive control on the prior or not, but the exact form in equation (3.11) suggests it will. Using the canonical form of the projection filter, given in equation (3.42), yields an evolution equation

$$\dot{\boldsymbol{\alpha}}_t + \dot{\boldsymbol{\beta}}_t = \mathbf{G}^{-1}(\boldsymbol{\alpha}_t + \boldsymbol{\beta}_t)\left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}(\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta},t)}[\boldsymbol{\phi}(\mathbf{x})] \right\rangle_{q(\mathbf{x},\boldsymbol{\alpha}_t,\boldsymbol{\beta}_t)}. \tag{6.15}$$

Constraining $\dot{\boldsymbol{\beta}} = 0$, as would be done in the forward pass of the discrete-time EP-smoothing algorithm, equation (6.15) yields a differential equation (in canonical space) for the forward message $\boldsymbol{\alpha}_t$. This could be integrated forward in time, in between observations, assuming the metric $\mathbf{G}(\boldsymbol{\alpha}_t + \boldsymbol{\beta}_t)$ can be dealt with, e.g. in the Gaussian setting. The expectations on the right-hand-side would be in moment form and would need to be mapped to canonical form using the link function. At an observation time, $\boldsymbol{\alpha}_t$ can be updated using the standard approach of EP. It is important to note that $q(\mathbf{x}, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$ is not the marginal density generated by the drift $\mathbf{g}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, t)$. A backward pass for updating $\boldsymbol{\beta}_t$ should involve the Kolmogorov forward operator $\overrightarrow{\mathcal{K}}_{\mathbf{g}(\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta},t)}$. Indeed, the *true* backward moment equations can be derived in terms of the backward operator and the likelihood $\psi(\mathbf{x}, t)$ (appendix A.3.2). There does not exist a *backward* projection filter in the literature (Brigo, 2011), but it should be possible to extend the geometric argument of section 3.4.2 to a projected version of the backward Kolmogorov equation. An alternative approach for the *Gaussian* EP setting, is to apply the Gaussian EP-smoothing algorithm to a Euler-Maruyama approximation and to take the limit as the time-step tends to zero. The conditionally Gaussian relations simplify many of the EP updates. Though as yet this approach has proved inconclusive.

6.2 Future work and extensions

The main direction for future work is experimental evidence. The topic closest to experimental implementation is the skew-GP smoother which simply requires the tracking of the diagonal of the third order moment. The use of skew-normal approximations in variational smoothing algorithm requires more preparation, which some difficult derivative-expectation relations that need to be finalised. For additive control drifts, it was already shown how the variational GP algorithm uses such a scheme. Future work will look at making this connection explicit, and also applying the variational GP method to more commonly well known stochastic optimal control problems.

6.2.1 Paper proposals

I currently see three future conference papers in the contents of this thesis.

- *Generalised variational smoothing:* Introduces the general variational smoothing framework of chapter 3, introduces the additional projection-filter constraints of chapter 3.4, introduces the gradient systems of chapter 4.

- *Variational skew GP smoothing:* Briefly introduces all of the above, then focuses on the skew-GP method of chapter 4.3.1, discusses the computations and results of the chapter in detail.
- *Variational GP for optimal control:* Focuses on the variational GP method with a rigorous exposition of its connection to approximate control. Discusses how the ideas can extended to more general distributions.

6.2.2 Dealing with high dimensions

One of the main difficulties with inference in partially observed processes, is dealing with high-dimensional hidden states. This is the spatial dimension of the state-space occupied by the state at any particular time, not the infinite dimension of the path-space. There are various approaches to dealing with these high dimensions, most of which can be borrowed from other problem settings. As with all approximate inference implementation problems, high dimensions can be dealt with in one of two ways, namely, adjusting the prior or adjusting the approximation. One of the earliest approaches to dealing with high dimensions in diffusion processes reduces the number of variables in the Fokker-Planck equation. The method proposed in (Risken, 1996, Section 8.3) is referred to as *adiabatic elimination* of the fast variables. In short, the dynamics of any system around an equilibrium will consist of variables with differing decay rates. The "fast" variables decay down to small values quickly and therefore an approximate system can ignore these variables and lower its dimension while still being able to represent the long term behaviour of the system. This is equivalent to a centre manifold type reduction of a stochastic system (Schumaker, 1987). To do this explicitly, i.e. reduce a stochastic system to its *normal form*, is at worst impossible and at best requires a lot of work to deal with eigenspace transformations and the technicalities of mapping noise variables through nonlinear mappings. But an approximate inference algorithm may not require such explicit formulations.

6.2.3 Hybrid methods

There is a current feeling that hybrid methods, combining sampling-based and deterministic methods, are the way forward for approximate inference in nonlinear state-space models (Murray & Storkey, 2011, Shen et al., 2010). One proposal is to use the variational approximation as the proposal distribution for a sampling-based method (Shen et al., 2010), but there are many other possibilities. Identifying the strong points of each of the different approximate inference methods is a key concern. Combining approximate inference methods and utilising these strong points is a natural extension to the set of current disjoint approaches. Variational methods are globally optimal, but as a result can lead to highly nonlinear optimisation problems. Therefore it seems natural to initialise variational methods in some way using EP or Monte Carlo methods. The hope would be to find global regions of low free-energy either quickly or accurately using a suboptimal method, and then to finesse solutions using variational methods to seek regions of local optimality. It should also be possible to establish a pre-processing stage to deterministic approximate inference where a principled analysis of the model is used to locate an appropriate class of approximation for the desired balance of accuracy and efficiency.

6.3 Final word

In all of science, engineering and technology, researchers and practitioners are frequently confronted with the problem of assimilating data into complex mathematical models encoding their prior beliefs. A coherent mathematical and algorithmic framework for blending dynamic models with time-series data has far reaching applications in fields such as neuroscience, climatology, econometrics, robotics, and bioinformatics. In this thesis, the variational smoothing framework of Archambeau & Opper (2011) was generalised to general drift-marginal pairings. The strict requirement of an explicit drift-marginal pairing was weakened by allowing marginal densities that drift into intractable regions to be projected back onto a specified class. These projected marginal densities can be used as surrogates for computing marginal density expectations required in the variational smoothing algorithm by gradient based optimisation methods. The approximate marginals were incorporated through continuous-time assumed-density type moment evolution constraints. Various novel variational approximations have been proposed, and the connections of the method to optimal control, kernel methods, and GP methods have been discussed. On the implementation side, a novel way for extending normal expectations of non-linear transformations to skew normal expectations, and the use of the true Hessian in gradient optimisations has been proposed.

Appendix A

Useful results

A.1 Finite-dimensional exponential families

Let $\mathbf{x}$ denote a random vector taking values in some space $\mathcal{X}$. Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_M)$ denote a collection of M real-valued functions $\phi_i : \mathcal{X} \to \mathbb{R}$, known as *sufficient statistics*. For a given vector of sufficient statistics $\boldsymbol{\phi}$, let $\boldsymbol{\theta}$ denote an associated vector of *canonical* or *exponential* parameters. For any realisation of $\mathbf{x}$, let $\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$ denote the Euclidean inner product in $\mathbb{R}^M$ of the two vectors $\boldsymbol{\theta}$ and $\boldsymbol{\phi}(\mathbf{x})$. Using this notation, the *exponential family* associated with $\boldsymbol{\phi}$ consists of the following parameterised collection of density functions

$$p(\mathbf{x}, \boldsymbol{\theta}) = \exp\Big(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})\Big), \tag{A.1}$$

where A denotes the *log partition* function

$$A(\boldsymbol{\theta}) = \log \int_{\mathcal{X}} \exp\Big(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\Big) d\mathbf{x}. \tag{A.2}$$

It is assumed that $\mathcal{X}$ is Lebesgue measurable (Dudley, 1993) with volume $d\mathbf{x}$. It is also assumed $A(\boldsymbol{\theta})$ is finite, and therefore $p(\mathbf{x}, \boldsymbol{\theta})$ is properly normalised. This occurs for all canonical parameters $\boldsymbol{\theta}$ belonging to the set

$$\Omega := \{\boldsymbol{\theta} \in \mathbb{R}^M | A(\boldsymbol{\theta}) < +\infty\}. \tag{A.3}$$

For more general domains and base measures $\nu(d\mathbf{x})$, and for additional source material, see Wainwright & Jordan (2008). Any exponential family has an alternative parameterisation in terms of a vector of *mean parameters*. More generally, for any density $p(\mathbf{x})$, the vector of mean parameters $\boldsymbol{\gamma}$ associated with the vector of statistics $\boldsymbol{\phi}(\mathbf{x})$ is defined by

$$\boldsymbol{\gamma} = \big\langle \boldsymbol{\phi}(\mathbf{x}) \big\rangle_p = \int \boldsymbol{\phi}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{A.4}$$

For any vector of statistics $\boldsymbol{\phi}(\mathbf{x})$, let $\mathcal{M}$ denote the set of realisable mean parameters

$$\mathcal{M} := \{\boldsymbol{\gamma} \in \mathbb{R}^M | \exists\, p \text{ s.t. } \left\langle \boldsymbol{\phi}(\mathbf{x}) \right\rangle_p = \boldsymbol{\gamma}\}. \tag{A.5}$$

The set $\mathcal{M}$ is not restricted to exponential family densities with $\boldsymbol{\phi}(\mathbf{x})$ as their vector of sufficient statistic. The set $\mathcal{M}$ is always a convex subset of $\mathbb{R}^M$.

A.1.1 Forward and backward mappings

For any exponential family density with vector of sufficient statistics $\boldsymbol{\phi}(\mathbf{x})$, there exists a *forward mapping* from the vector of canonical parameters $\boldsymbol{\theta} \in \Omega$ to the vector of mean parameters $\boldsymbol{\gamma} \in \mathcal{M}$, and a *backward mapping* from the vector of mean parameters $\boldsymbol{\gamma} \in \mathcal{M}$ to the vector of canonical parameters $\boldsymbol{\theta} \in \Omega$. Computation of either mapping can be extremely difficult and computationally intensive, especially in high dimensions. Let A^* denote the *conjugate function* of A, defined by

$$A^*(\boldsymbol{\gamma}) = \sup_{\boldsymbol{\theta} \in \Omega} \{\langle \boldsymbol{\gamma}, \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})\}. \tag{A.6}$$

Both $A(\boldsymbol{\theta})$ and $A^*(\boldsymbol{\gamma})$ are convex, and strictly convex if $\boldsymbol{\phi}(\mathbf{x})$ is *minimal*. Therefore A and A^* are *conjugate dual*, and A can be written

$$A(\boldsymbol{\theta}) = \sup_{\boldsymbol{\gamma} \in \mathcal{M}} \{\langle \boldsymbol{\theta}, \boldsymbol{\gamma} \rangle - A^*(\boldsymbol{\gamma})\}. \tag{A.7}$$

Equations (A.6) and (A.7) lead to the following result: For any exponential family density with vector of sufficient statistics $\boldsymbol{\phi}(\mathbf{x})$, the vector of canonical parameters $\boldsymbol{\theta} \in \Omega$ an the vector of mean parameters $\boldsymbol{\gamma} \in \mathcal{M}$ are related according to the dual matching conditions

$$\boldsymbol{\gamma}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}), \qquad \boldsymbol{\theta}(\boldsymbol{\gamma}) = \nabla_{\boldsymbol{\gamma}} A^*(\boldsymbol{\gamma}). \tag{A.8}$$

In general, $A(\boldsymbol{\theta})$ and $A^*(\boldsymbol{\gamma})$ do not have explicit forms and the Legendre pair $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ need to be found in tandem using approximate inference. Assume that $\boldsymbol{\gamma}$ lies in the interior of $\mathcal{M}$ and let $\boldsymbol{\theta}(\boldsymbol{\gamma})$ denote its dual under the matching condition in (A.8), then $A^*(\boldsymbol{\gamma})$ is given by a negative-entropy function

$$A^*(\boldsymbol{\gamma}) = -H(p(\mathbf{x}, \boldsymbol{\theta}(\boldsymbol{\gamma})) := \int_{\mathcal{X}} p(\mathbf{x}, \boldsymbol{\theta}(\boldsymbol{\gamma})) \log p(\mathbf{x}, \boldsymbol{\theta}(\boldsymbol{\gamma})) d\mathbf{x}. \tag{A.9}$$

This is not the standard definition of entropy because it is an extended real-valued function of $\boldsymbol{\gamma}$, rather than a functional of the density $p(\mathbf{x})$. To compute $A^*(\boldsymbol{\gamma})$ requires a sequence of mappings $\boldsymbol{\gamma} \mapsto \boldsymbol{\theta}(\boldsymbol{\gamma}) \mapsto p(\mathbf{x}, \boldsymbol{\theta}(\boldsymbol{\gamma})) \mapsto A^*(\boldsymbol{\gamma})$, see (Wainwright & Jordan, 2008, Section 3) for more details. For ease-of-use later, the forward and backward mappings will be written in compact form as an invertible *link* function $\boldsymbol{\omega} : \Omega \rightarrow \mathcal{M}$, such that

$$\begin{aligned} \boldsymbol{\omega}(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) && \text{(A.10)} \\ \boldsymbol{\omega}^{-1}(\boldsymbol{\gamma}) &= \nabla_{\boldsymbol{\gamma}} A^*(\boldsymbol{\gamma}). && \text{(A.11)} \end{aligned}$$

While it is relatively easy to construct an exponential family density with canonical parameter $\boldsymbol{\theta} \in \Omega$, it can be difficult or even impossible to compute its log partition function $A(\boldsymbol{\theta})$. The mean parameters $\boldsymbol{\gamma}(\boldsymbol{\theta}) \in \mathcal{M}$ are also often the quantities of interest, with low order moments of densities playing important roles in predictive learning problems. While no assumption has been made about these densities being Bayesian posteriors, realisations of data are easily absorbed into canonical parameters and sufficient statistics, and all of the above ideas are applicable to exponential family Bayesian posteriors. Captured in equation (A.7) is the understanding that both $A(\boldsymbol{\theta})$ and $\boldsymbol{\gamma}(\boldsymbol{\theta})$ can be found in tandem using a variational scheme. Additionally, phrasing inference as an optimisation problem provides a principled criteria of the *sub-optimality* of approximations. The importance of exponential families to learning in state-space models is how the canonical parameters and sufficient statistics allow for a natural partition of the approximating model into time-varying parameters and spatial statistics. By equipping $\boldsymbol{\theta}(t)$ with time dependencies, approximating densities can be tracked as they travel throughout the exponential family manifold, tied done to a particular family by the use of a time-invariant sufficient statistic $\boldsymbol{\phi}(\mathbf{x})$.

A.2 Fréchet derivative

To derive the exact continuous-discrete Bayesian smoothing solution in a variational formulation requires us to handle optimisation problems over infinite dimensional spaces. This type of problem is more inline with the traditional *calculus of variations* approach in mechanics (Fomin & Gelfand, 1963). The contents of this section is taken from (Fomin & Gelfand, 1963) and Cheney (2001). We require a *normed linear space*. An exact definition can be found in any classical real analysis book, e.g. Kolmogorov & Fomin (1975) or Folland (1984). The properties that are important to us are, if $\mathscr{E}$ is a normed linear space, then $\mathscr{E}$ has a norm $||\cdot||$, a zero element $0 \in \mathscr{E}$, and it is possible to define *linear functionals* on such as space.

Definition A.1. *Let $F : \Omega \rightarrow \mathbb{R}$ be a functional on an open set Ω in a normed linear space $\mathscr{E}$. For $f \in \Omega$, assume there exists a bounded linear map $A : \mathscr{E} \rightarrow \mathbb{R}$ such that*

$$\lim_{h \rightarrow 0} \frac{|F(f+h) - F(f) - A(h)|}{||h||} = 0. \tag{A.12}$$

Then F is said to be (Fréchet) differentiable *at f. Furthermore, A is called the* derivative *of F at f and is denoted $F'(f)$.*

It turns out that we will only require functionals quadratic and linear in their arguments, therefore the following examples focus on these two cases.

Example A.1 (Abstract, bounded linear functional)**.** *Let F be a bounded linear functional on $\mathscr{E}$. Then $F'(f) = F(\cdot)$ follows from the relation*

$$|F(f+h) - F(f) - A(h)| = |F(h) - A(h)| = 0. \tag{A.13}$$

Example A.2 (Hilbert, bounded linear functional)**.** *Let $\mathscr{E} = \mathcal{H}$ denote a Hilbert space with inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$. For any $q \in \mathcal{H}$, define $F_q(\cdot) = \langle q, \cdot \rangle$. From the* Riesz representation theorem *(Folland, 1984), any linear functional $A(\cdot)$ on $\mathcal{H}$ can be written $A_g(\cdot) = \langle g, \cdot \rangle$ for some $g \in \mathcal{H}$. Then $F'(f) = F_q(\cdot)$ follows from the relation*

$$|F_q(f+h) - F_q(f) - A_g(h)| = |\langle q - g, h \rangle| = 0. \tag{A.14}$$

Example A.3 (Quadratic form)**.** *Let $\mathscr{E} = \mathcal{H}$ denote a Hilbert space with inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$. For a bounded linear operator $L : \mathcal{H} \to \mathcal{H}$, define $F_L(\cdot) = \langle \cdot, L \cdot \rangle$. Then for any $f \in \mathcal{H}$, it holds that $F'(f) = A_{L,f}(\cdot)$ where*

$$A_{L,f}(h) = \langle L^* f + Lf, h \rangle. \tag{A.15}$$

Example A.4 (Maximum entropy)**.** *Let $V(\mathbf{x}, t)$ denote the energy of a spatio-temporal system. Let $q(\mathbf{x}, t)$ denote the variational marginal density at time t. Let $\lambda \in L^2([0,T])$ denote the Lagrange multiplier for the constraint $\int q(\mathbf{x}, t) d\mathbf{x} = 1$ for all $t \in [0,T]$. Then the corresponding* Maximum entropy Lagrangian *$\mathcal{L}(q, \lambda)$ for $q \in L^2(\mathbb{R}^{d_{\mathbf{x}}} \times [0,T])$ is given by*

$$\mathcal{L}(q, \lambda) \;\; = \;\; \int_0^T \Big(\big\langle V(\mathbf{x}, t) + \log q(\mathbf{x}, t) + \lambda(t) \big\rangle_{q(\mathbf{x},t)} - \lambda(t) \Big) dt, \tag{A.16}$$

where it is assumed that $-\langle \log q \rangle_q < \infty$. Equating $\partial_q \mathcal{L}(q, \lambda) = 0$ yields

$$q(\mathbf{x}, t) = \exp\Big(- V(\mathbf{x}, t) + 1 + \lambda(t) \Big) \tag{A.17}$$

where $\lambda(t)$ is chosen so that $\int q(\mathbf{x}, t) d\mathbf{x} = 1$ for all $t \in [0,T]$.

A.3 Supplementary proofs

A.3.1 Posterior Wiener process

Though the focus of this thesis in on general nonlinear relations, some fundamental properties of the model are captured by considering general linear relations. The humble Gaussian appears all throughout this thesis in different forms - from static univariate Gaussians, to discrete-time jointly Gaussian filters, to Gaussian measures over continuous-time sample paths. Inference in all these settings follows the same simple rule for conditioning a Gaussian random variable, given below in its most general form (Prato & Zabczyk, 1992).

Theorem A.2. *Let $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ denote a separable Hilbert space. Let $(u_1, u_2) \in \mathcal{H}_1 \oplus \mathcal{H}_2$ denote an $\mathcal{H}$-valued Gaussian random variable with mean $m = (m_1, m_2)$ and covariance operator $\mathcal{C} \in (\mathcal{H}_1 \oplus \mathcal{H}_2) \otimes (\mathcal{H}_1 \oplus \mathcal{H}_2)$, such that $\mathcal{C}_{ij} = \big\langle (u_i - m_i) \otimes (u_j - m_j) \big\rangle$ for $i, j = 1, 2$. Then the conditional distribution of u_1 given u_2 is Gaussian with mean m' and covariance operator $\mathcal{C}'$ given by*

$$m' = m_1 + \mathcal{C}_{12}\mathcal{C}_{22}^{-1}(u_2 - m_2), \qquad \mathcal{C}' = \mathcal{C}_{11} - \mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{21}. \tag{A.18}$$

The generality of theorem A.2 allows almost all linear inference problems to be dealt with in a universal fashion. For the example in section 2.2.2.1, assume $x(\cdot)$ is a random variable taking values in the space of square integrable functions $L^2([0, T])$. The corresponding prior covariance operator $\Lambda \in L^2([0, T]) \otimes L^2([0, T])$ is defined using the integral equation

$$\langle \phi, \Lambda\varphi \rangle_{L^2} = \langle \Lambda\phi, \varphi \rangle_{L^2} = \int\int k(\tau, t)\phi(\tau)\varphi(t) d\tau dt \tag{A.19}$$

for any $\phi, \varphi \in L^2([0, T])$. The observation model can be written in general form

$$y = \langle x, \delta_s \rangle_{L^2} + \eta \tag{A.20}$$

where $\langle \cdot, \cdot \rangle_{L^2}$ is the inner product in $L^2([0, T])$ and $\delta_s(\cdot)$ is the evaluation mapping centered at s. Tailoring the model to theorem A.2, the example consists of a joint Gaussian measure over $(x, y) \in L^2([0, T]) \oplus \mathbb{R}$ with mean function $m \in L^2([0, T]) \oplus \mathbb{R}$ and covariance operator $\mathcal{C} \in (L^2([0, T]) \oplus \mathbb{R}) \otimes (L^2([0, T]) \oplus \mathbb{R})$ given by

$$m = (0, 0), \qquad \langle (\phi, a), \mathcal{C}(\varphi, b) \rangle_{L^2 \oplus \mathbb{R}} = \begin{pmatrix} \langle \phi, \Lambda\varphi \rangle_{L^2} & \langle \Lambda\delta_s, \varphi \rangle_{L^2} \\ \langle \phi, \Lambda\delta_s \rangle_{L^2} & ab(\langle \Lambda\delta_s, \delta_s \rangle_{L^2} + r) \end{pmatrix} \tag{A.21}$$

for any $(\phi, a), (\varphi, b) \in L^2([0, T]) \oplus \mathbb{R}$. Inserting these equations into theorem (A.2) leads directly to a posterior $\mathcal{N}(\mu, \Sigma)$ with mean $\mu \in L^2([0, T])$ and covariance operator $\Sigma \in L^2([0, T]) \otimes L^2([0, T])$ given by

$$\mu = \frac{y\Lambda\delta_s}{\langle \Lambda\delta_s, \delta_s \rangle_{L^2} + r}, \qquad \langle \phi, \Sigma\varphi \rangle_{L^2} = \langle \phi, \Lambda\varphi \rangle_{L^2} - \frac{\langle \Lambda\delta_s, \varphi \rangle_{L^2} \langle \phi, \Lambda\delta_s \rangle_{L^2}}{\langle \Lambda\delta_s, \delta_s \rangle_{L^2} + r}. \tag{A.22}$$

The posterior mean and covariance can be projected onto a particular time $t \in [0, T]$ to give a mean and variance value for the hidden state $x(t)$. Indeed, for any $t, \tau \in [0, T]$ it holds that

$$\mu(t) := \langle \mu, \delta_s \rangle_{L^2} = \frac{yk(s, t)}{k(s, s) + r}, \qquad \kappa(t, \tau) := \langle \delta_t, \Sigma\delta_\tau \rangle_{L^2} = k(t, \tau) - \frac{k(s, t)k(\tau, s)}{k(s, s) + r}. \tag{A.23}$$

A.3.2 Backward Kolmogorov moment equations

For any twice-differentiable scalar function $\phi(\mathbf{x})$, the integral $\langle\phi(\mathbf{x})\rangle_{\psi(\mathbf{x},t)}$ satisfies the differential equation

$$\frac{\partial\langle\phi(\mathbf{x})\rangle_{\psi(\mathbf{x},t)}}{\partial t} = -\left\langle\overrightarrow{\mathcal{K}}_{\mathbf{g}}[\phi(\mathbf{x})]\right\rangle_{\psi(\mathbf{x},t)} \tag{A.24}$$

where $\overrightarrow{\mathcal{K}}_{\mathbf{g}}$ is the forward operator defined in (2.23). This follows directly from the backward equation and integration by parts.

Proof. Assuming differentiability, we have

$$\begin{aligned}\frac{\partial\langle\phi(\mathbf{x})\rangle_{\psi(\mathbf{x},t)}}{\partial t} &= \int\phi(\mathbf{x})\frac{\partial\psi(\mathbf{x},t)}{\partial t}d\mathbf{x} && \text{(A.25)}\\ &= -\int\phi(\mathbf{x})\overleftarrow{\mathcal{K}}[\psi(\mathbf{x},t)]d\mathbf{x} && \text{(A.26)}\\ &= -\int\psi(\mathbf{x},t)\overrightarrow{\mathcal{K}}[\phi(\mathbf{x})]d\mathbf{x}. && \text{(A.27)}\end{aligned}$$

□

Therefore it is possible to obtain propagation and update equations for any finite moment of $\psi(\mathbf{x},t)$, while (A.24) requires full knowledge of all other moments.

A.3.3 Gaussian expectation-derivative relations

Lemma A.3. *For any two univariate-normal random variables z and x, with respective means m and n, and any smooth function $f:\mathbb{R}\to\mathbb{R}$, it holds that*

$$\langle zf(x)\rangle = m\langle f(z)\rangle + \mathrm{Cov}(z,x)\partial_n\langle f(x)\rangle. \tag{A.28}$$

Lemma A.3. The characteristic function of a d-dimensional multivariate normal random vector $\mathbf{z}=(z_1,\ldots,z_d)$ with mean $\mathbf{m}\in\mathbb{R}^d$ and covariance $\mathbf{S}\in\mathbb{R}^{d\times d}$ is given by

$$\phi(\boldsymbol{\omega}) := \exp\Big((-1)^{1/2}\boldsymbol{\omega}^\top\big(\mathbf{m}-\tfrac{1}{2}\mathbf{S}\boldsymbol{\omega}\big)\Big). \tag{A.29}$$

For $\{i_k\in\{1,\ldots d\}\}_{k=1}^N$, $N\in\mathbb{N}$, it holds that

$$\langle z_{i_1}\cdots z_{i_N}\rangle = (-1)^{-N/2}\partial_{\omega_{i_1}}\cdots\partial_{\omega_{i_N}}\phi(\boldsymbol{\omega})\Big|_{\boldsymbol{\omega}=0}. \tag{A.30}$$

Define $C_i(\boldsymbol{\omega}) := \partial_{\omega_i}\phi(\boldsymbol{\omega}) = (-1)^{1/2}m_i - \sum_{j=1}^{d}\omega_j S_{j,i}$, and let $\psi_k : \mathbb{C} \to \mathbb{C}$ be the function such that $\partial_{\omega_i}^k \phi(\boldsymbol{\omega}) = \psi_k\big(C_i(\boldsymbol{\omega})\big)\phi(\boldsymbol{\omega})$. From (A.30) we have

$$\begin{aligned} \langle z_i^k\rangle &= (-1)^{-N/2}\psi_k\big(C_i(0)\big)\phi(0) && \text{(A.31)}\\ &= (-1)^{-N/2}\psi_k\big((-1)^{1/2}m_i\big) && \text{(A.32)}\\ \Rightarrow \quad \partial_{m_i}\langle z_i^k\rangle &= (-1)^{-N/2}\partial_{m_i}\psi_k\big((-1)^{1/2}m_i\big) && \text{(A.33)}\\ &= (-1)^{-\frac{N-1}{2}}\psi_k'\big((-1)^{1/2}m_i\big). && \text{(A.34)} \end{aligned}$$

Therefore, inserting (A.32) and (A.34)

$$\begin{aligned} \langle z_j z_i^k\rangle &= (-1)^{-\frac{N+1}{2}}\partial_{\omega_j}\psi_k\big(C_i(\boldsymbol{\omega})\big)\phi(\boldsymbol{\omega})\Big|_{\boldsymbol{\omega}=0} && \text{(A.35)}\\ &= (-1)^{-\frac{N+1}{2}}\Big[C_j(\boldsymbol{\omega})\psi_k\big(C_i(\boldsymbol{\omega})\big) - S_{i,j}\psi_k'\big(C_i(\boldsymbol{\omega})\big)\Big]\phi(\boldsymbol{\omega})\Big|_{\boldsymbol{\omega}=0} && \text{(A.36)}\\ &= (-1)^{-\frac{N+1}{2}}\Big[(-1)^{1/2}m_j\psi_k\big((-1)^{1/2}m_i\big) - S_{i,j}\psi_k'\big((-1)^{1/2}m_i\big)\Big] && \text{(A.37)}\\ &= m_j\langle z_i^k\rangle + S_{i,j}\partial_{m_i}\langle z_i^k\rangle. && \text{(A.38)} \end{aligned}$$

Taking the Taylor series of $f(x)$ around 0, we have

$$\begin{aligned} \langle z_j f(z_i)\rangle &= \sum_{k=1}^{\infty} a_i\langle z_j z_i^k\rangle && \text{(A.39)}\\ &= \sum_{k=1}^{\infty} a_i\big(m_j\langle z_i^k\rangle + S_{i,j}\partial_{m_i}\langle z_i^k\rangle\big) && \text{(A.40)}\\ &= m_j\langle f(z_i)\rangle + S_{i,j}\partial_{m_i}\langle f(z_i)\rangle. && \text{(A.41)} \end{aligned}$$

□

Using the fact that $\partial_n\langle f(x)\rangle = \langle\partial_x f(x)\rangle$ and applying the lemma to each dimension individually, we obtain

$$\langle f^{\mathrm{T}}(\mathbf{x})\mathbf{x}\rangle = \langle f^{\mathrm{T}}(\mathbf{x})\boldsymbol{\mu}\rangle + \langle\nabla_{\mathbf{x}} f(\mathbf{x})\rangle\boldsymbol{\Sigma}. \tag{A.42}$$

A.3.4 Free-energy for second-order motions

Define $\mathbf{z} = (\mathbf{x}, \mathbf{v})$. For $0 = t_0 < t_1 < \cdots < t_K = T$, $t_{k+1} - t_k = \Delta t$, let $\mathbf{z}^{(0:K)} = (\mathbf{z}^{(0)}, \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(K)})$ be the discrete time path with $\mathbf{z}^{(k)} = \mathbf{z}(t_k)$ and let $\mathbf{z}(0)$ be known. Using the Markov property

$$\begin{aligned} p(\mathbf{z}^{(1:K)}) &= \prod_{k=0}^{K-1} p(\mathbf{x}^{(k+1)}|\mathbf{z}^{(k)})p(\mathbf{v}^{(k+1)}|\mathbf{z}^{(k)})\\ q(\mathbf{z}^{(1:K)}) &= \prod_{k=0}^{K-1} q(\mathbf{x}^{(k+1)}|\mathbf{z}^{(k)})q(\mathbf{v}^{(k+1)}|\mathbf{z}^{(k)}) \end{aligned}$$

where, for $\delta(\cdot)$ the indicator function at the origin,

$$\begin{aligned} p(\mathbf{x}^{(k+1)}|\mathbf{z}^{(k)}) &= \delta(\mathbf{x}^{(k)} + \mathbf{v}^{(k)}\Delta t - \mathbf{x}^{(k+1)}) \\ p(\mathbf{v}^{(k+1)}|\mathbf{z}^{(k)}) &= \mathcal{N}(\mathbf{v}^{(k+1)}|\mathbf{v}^{(k)} + \mathbf{f}(t_k, \mathbf{z}^{(k)})\Delta t, \Sigma\Delta t) \end{aligned}$$

and

$$\begin{aligned} q(\mathbf{x}^{(k+1)}|\mathbf{z}^{(k)}) &= \delta(\mathbf{x}^{(k)} + \mathbf{v}^{(k)}\Delta t - \mathbf{x}^{(k+1)}) \\ q(\mathbf{v}^{(k+1)}|\mathbf{z}^{(k)}) &= \mathcal{N}(\mathbf{v}^{(k+1)}|\mathbf{v}^{(k)} + \mathbf{g}(t_k, \mathbf{z}^{(k)})\Delta t, \Sigma\Delta t). \end{aligned}$$

The relative entropy $\widetilde{\mathrm{KL}}$ for the discretised problem is given by

$$\begin{aligned} \widetilde{\mathrm{KL}} &:= \mathrm{KL}[q(\mathbf{z}^{(1:K)})||p(\mathbf{z}^{(1:K)})] \\ &= \sum_{k=0}^{K-1} \left\langle \left\langle \underbrace{\ln \frac{q(\mathbf{x}^{(k+1)}|\mathbf{z}^{(k)})}{p(\mathbf{x}^{(k+1)}|\mathbf{z}^{(k)})}}_{=0} + \ln \frac{q(\mathbf{v}^{(k+1)}|\mathbf{z}^{(k)})}{p(\mathbf{v}^{(k+1)}|\mathbf{z}^{(k)})} \right\rangle_{q(\mathbf{z}^{(k+1)}|\mathbf{z}^{(k)})} \right\rangle_{q(\mathbf{z}^{(k)})} \\ &= \frac{\Delta t}{2} \sum_{k=0}^{K-1} \langle ||\mathbf{f}(t_k, \mathbf{z}^{(k)}) - \mathbf{g}(t_k, \mathbf{z}^{(k)}))||_{\Sigma}^2 \rangle_{q(\mathbf{z}^{(k)})}. \end{aligned}$$

The claim follows by taking the limit as $\Delta t \to 0$.

A.3.5 Third-order moment evolution equations

Proof of (4.39), (4.40)*, and* (4.41). To ease notation, define $\mathbf{g} := \mathbf{g}(\mathbf{x}, t)$ and $\mathbf{D} := \mathbf{D}(\mathbf{x}, t)$. Inserting $\phi(\mathbf{x}) = x^{(i)}$ into (3.13) gives

$$\begin{aligned} \dot{m}_1^{(i)} &= \left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}[x^{(i)}] \right\rangle_{q_t} && \text{(A.43)} \\ &= \left\langle \mathbf{g}^{\mathrm{T}} \nabla x^{(i)} + \tfrac{1}{2}\mathrm{tr}\big(\mathbf{D}(\nabla\nabla^{\mathrm{T}})x^{(i)}\big) \right\rangle_{q_t} && \text{(A.44)} \\ &= \langle g^{(i)} \rangle_{q_t}. && \text{(A.45)} \end{aligned}$$

Performing (A.45) for each $i \in \mathbb{N}_{d_x}$ gives (4.39). Inserting $\phi(\mathbf{x}) = \big(x^{(i)} - m_1^{(i)}\big)\big(x^{(i)} - m_1^{(i)}\big)$ into (3.13) gives

$$\begin{aligned} \dot{m}_2^{(ij)} &= \left\langle \overleftarrow{\mathcal{K}}_{\mathbf{g}}\Big[\big(x^{(i)} - m_1^{(i)}\big)\big(x^{(i)} - m_1^{(i)}\big)\Big] \right\rangle_{q_t} && \text{(A.46)} \\ &= \left\langle \mathbf{g}^{\mathrm{T}}\Big(\big(x^{(j)} - m_1^{(j)}\big)\mathbf{e}_i + \big(x^{(i)} - m_1^{(i)}\big)\mathbf{e}_j\Big) + \tfrac{1}{2}\mathrm{tr}\big(\mathbf{D}(\mathbf{e}_j\mathbf{e}_i^{\mathrm{T}} + \mathbf{e}_i\mathbf{e}_j^{\mathrm{T}})\big) \right\rangle_{q_t} && \text{(A.47)} \\ &= \left\langle g^{(i)}\big(x^{(j)} - m_1^{(j)}\big) + g^{(j)}\big(x^{(i)} - m_1^{(i)}\big) + D^{(ij)} \right\rangle_{q_t}, && \text{(A.48)} \end{aligned}$$

where we have used

$$\begin{aligned}\nabla\big(x^{(i)}-m_1^{(i)}\big)\big(x^{(j)}-m_1^{(j)}\big) &= \big(x^{(j)}-m_1^{(j)}\big)\mathbf{e}_i+\big(x^{(i)}-m_1^{(i)}\big)\mathbf{e}_j &\text{(A.49)}\\ (\nabla\nabla^{\mathrm{T}})\big(x^{(i)}-m_1^{(i)}\big)\big(x^{(j)}-m_1^{(j)}\big) &= \mathbf{e}_j\mathbf{e}_i^{\mathrm{T}}+\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}}. &\text{(A.50)}\end{aligned}$$

Performing (A.48) for each $(i,j)\in\mathbb{N}_{d_x}\times\mathbb{N}_{d_x}$ gives (4.40). Finally, inserting $\phi(\mathbf{x})=\prod_{l\in\{i,j,k\}}\big(x^{(l)}-m_1^{(l)}\big)$ into (3.13) gives

$$\begin{aligned}\dot{m}_3^{(iii)} &= \Big\langle\overleftarrow{\mathcal{K}}_{\mathbf{g}}\big[\prod_{l\in\{i,j,k\}}\big(x^{(l)}-m_1^{(l)}\big)\big]\Big\rangle_{q_t} &\text{(A.51)}\\ &= \sum_{l\in\{i,j,k\}}\Big\langle g^{(l)}\prod_{n\in\{i,j,k\}/l}\big(x^{(n)}-m_1^{(n)}\big)\Big\rangle\\ &\quad+\sum_{l\in\{i,j,k\}}\sum_{\substack{n\in\{i,j,k\}/l\\ w\in\{i,j,k\}/\{l,n\}}}\Big\langle D^{(nl)}\big(x^{(w)}-m_1^{(w)}\big)\Big\rangle &\text{(A.52)}\end{aligned}$$

where we have used

$$\begin{aligned}\nabla\phi(\mathbf{x}) &= \nabla\prod_{l\in\{i,j,k\}}\big(x^{(l)}-m_1^{(l)}\big) &\text{(A.53)}\\ &= \sum_{l\in\{i,j,k\}}\prod_{n\in\{i,j,k\}/l}\big(x^{(n)}-m_1^{(n)}\big)\mathbf{e}_l &\text{(A.54)}\end{aligned}$$

and

$$\begin{aligned}\nabla\nabla\phi(\mathbf{x}) &= \nabla\sum_{l\in\{i,j,k\}}\prod_{n\in\{i,j,k\}/l}\big(x^{(n)}-m_1^{(n)}\big)\mathbf{e}_l &\text{(A.55)}\\ &= \sum_{l\in\{i,j,k\}}\sum_{\substack{n\in\{i,j,k\}/l\\ w\in\{i,j,k\}/\{l,n\}}}\big(x^{(w)}-m_1^{(w)}\big)\mathbf{e}_n\mathbf{e}_l^{\mathrm{T}}. &\text{(A.56)}\end{aligned}$$

□

A.3.6 Deriving GPs from linear state-space models

Consider the process $\{\mathbf{x}_t\}_{t\in T}$ whose law is described by the linear SDE in equation (5.14). Using Itô's formula, the explicit form for $\mathbf{x}_t$ is given by

$$\mathbf{x}_t=e^{\int_0^t\mathbf{A}(s)ds}\mathbf{x}_0+\int_0^t e^{\int_s^t\mathbf{A}(u)du}\mathbf{b}(s)ds+\int_0^t e^{\int_s^t\mathbf{A}(u)du}\sqrt{\mathbf{D}}_s d\mathbf{w}_s. \tag{A.57}$$

Therefore

$$\boldsymbol{\mu}(t):=\langle\mathbf{x}_t\rangle=e^{\int_0^t\mathbf{A}(s)ds}\boldsymbol{\mu}_0+\int_0^t e^{\int_s^t\mathbf{A}(u)du}\mathbf{b}(s)ds. \tag{A.58}$$

Also

$$\begin{aligned}\boldsymbol{\Sigma}(s,t) &:= \left\langle (\mathbf{x}_s - \boldsymbol{\mu}_s)(\mathbf{x}_t - \boldsymbol{\mu}_t)^{\mathrm{T}} \right\rangle && \text{(A.59)}\\ &= \left\langle \left(e^{\int_0^s \mathbf{A}(u)du}(\mathbf{x}_0 - \boldsymbol{\mu}_0) + \int_0^s e^{\int_u^s \mathbf{A}(v)dv}\sqrt{\mathbf{D}}_u d\mathbf{w}_u \right)\right.\\ &\qquad \left. \times \left(e^{\int_0^t \mathbf{A}(u)du}(\mathbf{x}_0 - \boldsymbol{\mu}_0) + \int_0^t e^{\int_u^t \mathbf{A}(v)dv}\sqrt{\mathbf{D}}_u d\mathbf{w}_u \right)^{\mathrm{T}} \right\rangle && \text{(A.60)}\\ &= e^{\int_0^s \mathbf{A}(u)du}\boldsymbol{\Sigma}_0 e^{\int_0^t \mathbf{A}^{\mathrm{T}}(u)du} + \int_0^{t\wedge s} e^{\int_u^s \mathbf{A}(v)dv}\mathbf{D}_u e^{\int_u^t \mathbf{A}^{\mathrm{T}}(v)dv} du, && \text{(A.61)}\end{aligned}$$

where the second term in (A.61) is derived using a variation of the Itô isometry (Øksendal, 2003).

A.3.7 Reproducing kernel of first-order Sobolev space

By definition, $x \in \mathcal{X}$ if and only if $x' + \lambda x \in L^2([0,T],\mathbb{R})$ and $x(0) \in \mathbb{R}$. Therefore $\mathcal{X}$ inherits a natural bilinear form $\langle \cdot,\cdot \rangle_\lambda : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, from the L^2 inner product $\langle \cdot,\cdot \rangle_{L^2}$ and the scalar product on $\mathbb{R}$, given by

$$\langle x,y \rangle_\lambda = \langle x' + \lambda x, y' + \lambda y \rangle_{L^2} + 2\lambda x(0)y(0). \tag{A.62}$$

The covariance function in equation (5.40) is non-differentiable, so we restate it in *twice-differentiable* piecewise form

$$k(t,s) = \begin{cases} l_t(s) = \frac{1}{2\lambda}e^{-\lambda(t-s)} & \text{for } s \in [0,t] \\ u_t(s) = \frac{1}{2\lambda}e^{-\lambda(s-t)} & \text{for } t \in [t,T]. \end{cases} \tag{A.63}$$

Using *integration-by-parts* twice on $[0,t]$ and $[t,T]$, for any $x \in \mathcal{X}$, assuming $k(t,\cdot) \in \mathcal{X}$, it holds that

$$\begin{aligned}\langle x, k(t,\cdot) \rangle_\lambda &= \int_0^t x(s)(\lambda^2 l_t(s) - l_t''(s))ds + [x(s)(\lambda l_t(s) + l_t'(s))]_0^t + 2\lambda x(0) l_t(0)\\ &\quad + \int_t^T x(s)(\lambda^2 u_t(s) - u_t''(s))ds + [x(s)(\lambda u_t(s) + u_t'(s))]_t^T. && \text{(A.64)}\end{aligned}$$

Inserting definitions (A.63) into (A.64), and evaluating the right-hand-side of (A.64), we obtain $\langle x, k(t,\cdot) \rangle_\lambda = x(t)$, which is the desired *reproducing* property. Note that $\langle \cdot,\cdot \rangle_\lambda$ is symmetric, and $\langle k(t,\cdot), k(t,\cdot) \rangle_\lambda = k(t,t)$ is finite so the assumption $k(x,\cdot) \in \mathcal{X}$ was correct. It can additionally be shown that $(\mathcal{X}, k)$ is a *unique* RKHS (Aronszajn, 1950) by noting that (5.40) is positive-definite for $\lambda > 0$. Interestingly, note the reproducing property did *not* require positive-definiteness (see Canu et al. (2002) for a discussion).

A.3.8 Cumulants of skew-normal distribution

For any random variable $\mathbf{x}$, let $\mathcal{K}(\mathbf{z})$ denote its cumulant generating function (CGF) given by

$$\mathcal{K}(\mathbf{z}) = \log \big\langle \exp(\mathbf{x}^{\mathrm{T}}\mathbf{z}) \big\rangle. \tag{A.65}$$

Let $\boldsymbol{\kappa}_1$, $\boldsymbol{\kappa}_2$, and $\boldsymbol{\kappa}_3$ denote the corresponding first three cumulants, given by

$$\begin{aligned} \boldsymbol{\kappa}_1 &:= \nabla_{\mathbf{z}}\mathcal{K}(\mathbf{0}) && \text{(A.66)} \\ \boldsymbol{\kappa}_2 &:= \nabla_{\mathbf{z}\mathbf{z}^{\mathrm{T}}}\mathcal{K}(\mathbf{0}) && \text{(A.67)} \\ \kappa_3^{(ijk)} &:= \partial_{z_i z_j z_k}\mathcal{K}(\mathbf{0}), \quad i,j,k \in \mathbb{N}_{d_{\mathbf{x}}}. && \text{(A.68)} \end{aligned}$$

The cumulant generating function $\mathcal{K}(\mathbf{z})$ of the skew-normal distribution in (ref) is given by

$$\mathcal{K}(\mathbf{z}) = \boldsymbol{\mu}^{\mathrm{T}}\mathbf{z} + \frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{z} + \zeta(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}) \tag{A.69}$$

where $\zeta(x) = \log\big(2\Phi(x)\big)$ and $\boldsymbol{\delta}$ is given by (ref). To take derivatives of $\mathcal{K}(\mathbf{z})$, and therefore find the cumulants, requires us to take derivatives of $\zeta(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z})$ w.r.t. $\mathbf{z}$. These are given by

$$\begin{aligned} \nabla_{\mathbf{z}}\zeta(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}) &= \zeta'(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z})\boldsymbol{\delta} && \text{(A.70)} \\ \nabla_{\mathbf{z}\mathbf{z}}\zeta(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}) &= \zeta''(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z})\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}} && \text{(A.71)} \\ \partial_{z_i z_j z_k}\zeta(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}) &= \zeta'''(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z})\delta_i\delta_j\delta_k. && \text{(A.72)} \end{aligned}$$

These need to be evaluated at $t = 0$, therefore we require the following identities (Azzalini & Capitanio, 1999)

$$\begin{aligned} \zeta'(0) &= (2/\pi)^{1/2} && \text{(A.73)} \\ \zeta''(0) &= (2/\pi) && \text{(A.74)} \\ \zeta'''(0) &= (2/\pi^3)^{1/2}(4-\pi). && \text{(A.75)} \end{aligned}$$

Putting it all together, we obtain

$$\begin{aligned} \boldsymbol{\kappa}_1 &:= \nabla_{\mathbf{z}}\mathcal{K}(\mathbf{0}) = \boldsymbol{\mu} + (2/\pi)^{1/2}\boldsymbol{\delta} && \text{(A.76)} \\ \boldsymbol{\kappa}_2 &:= \nabla_{\mathbf{z}\mathbf{z}}\mathcal{K}(\mathbf{0}) = \boldsymbol{\Lambda} - (2/\pi)\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}} && \text{(A.77)} \\ \kappa_3^{(ijk)} &:= \partial_{z_i z_j z_k}\mathcal{K}(\mathbf{0}) = (2/\pi^3)^{1/2}(4-\pi)\delta_i\delta_j\delta_k. && \text{(A.78)} \end{aligned}$$

A.4 Likelihood analysis of double well

The negative log likelihood is given by

$$\widehat{\mathcal{F}}(\mu; b, \theta, y) = \frac{(\mu - y)^2}{2} + b\mu^2(\mu^2 - 2\theta). \tag{A.79}$$

Therefore finding the critical points of $\widehat{\mathcal{F}}(\mu; b, \theta, y)$ is equivalent to finding the zeros of the cubic equation

$$\mu^3 + A\mu + B = 0 \tag{A.80}$$

where $A = (\frac{1}{4b} - \theta)$ and $B = \frac{-y}{4b}$. Using standard rules for cubic equations (Abramowitz & Stegun, 1964), equation (A.80) has (three) *real* roots if and only if $4A^3 + 27B^2 \leq 0$. This is equivalent to the identity

$$|y| \leq 8b\sqrt{\frac{(\theta - \frac{1}{4b})^3}{27}}. \tag{A.81}$$

For (A.81) to hold, it must already hold that $(\theta - \frac{1}{4b}) \geq 0$ and therefore $\theta \geq \frac{1\epsilon}{4r}$, where $b = \frac{\epsilon}{r}$ has been re-substituted momentarily for the benefit of discussion. Interpreting this qualitatively, consider the case when $\theta < \frac{1\epsilon}{4r}$, and thus $\widehat{\mathcal{F}}(\mu; b, \theta, y)$ has only one minima (due to global stability). There are two possible explanations why this type of behaviour occurs. If the observation noise r is small when compared to ϵ and θ, then it is possible that $\theta < \frac{1\epsilon}{4r}$. For this case the post hoc Laplace posterior is not able to "see" the minima of the prior potential $U(\mathbf{x})$ because the bandwidth of the likelihood function is too tight. The same situation occurs if ϵ is large when compared θ and r, and thus $\theta < \frac{1\epsilon}{4r}$. In this case the diffusion rate (temperature) of the prior is too high and, because of the level of noise in the system, it is impossible for the Laplace approximation to distinguish between the two minima of the prior potential $U(\mathbf{x})$. This type of behaviour can be seen in figure (..). Note that identity (A.81) is equivalent

$$-1 \leq \frac{y}{8b}\sqrt{\frac{27}{(\theta - \frac{1}{4b})^3}} \leq 1. \tag{A.82}$$

this is important because, for the case $\theta \geq \frac{1}{4b}$, equation (A.82) allows the global minimiser $\mu^*(b, \theta, y)$ of $\widehat{\mathcal{F}}(\mu; b, \theta, y)$ to be expressed in closed trigonometric form. Indeed, for the case $\theta \geq \frac{1}{4b}$, the minimiser $\mu^*(b, \theta, y)$ is given by

$$\mu^*(b,\theta,y) = \begin{cases} 2\sqrt{\frac{\theta - \frac{1}{4b}}{3}}\cos\Big(\frac{2\pi}{3} - \phi(b,\theta,y)\Big) & \text{if } 0 \geq y \\ 2\sqrt{\frac{\theta - \frac{1}{4b}}{3}}\cos\Big(\frac{2\pi}{3} + \phi(b,\theta,y)\Big) & \text{if } 0 \leq y \end{cases} \tag{A.83}$$

where

$$\phi(b,\theta,y) \quad = \quad \frac{1}{3}\cos^{-1}\left(\frac{y}{8b}\sqrt{\frac{27}{(\theta - \frac{1}{4b})^3}}\right). \tag{A.84}$$

$\widehat{\mathcal{F}}$ has first and second order derivatives given by

$$\begin{aligned} \widehat{\mathcal{F}}_\mu &= (\mu - y) + 4b\mu(\mu^2 - \theta) && \text{(A.85)} \\ \widehat{\mathcal{F}}_{\mu\mu} &= 1 + 4b(3\mu^2 - \theta). && \text{(A.86)} \end{aligned}$$

Setting $\widehat{\mathcal{F}}_{\mu\mu} = 0$ shows two inflection points exist in the graph of $\widehat{\mathcal{F}}(\mu)$, and their positions are independent of y. Indeed, the inflection points are given by

$$\mu = \pm\sqrt{\tfrac{1}{3}(\theta - \tfrac{1\epsilon}{4r})} \tag{A.87}$$

It follows that these inflection points only exist for $\theta > \frac{1\epsilon}{4r}$. Now assume that $\theta > \frac{1}{4b}$, and therefore two inflection points and importantly three minima exist in the graph of $\widehat{\mathcal{F}}(\mu)$. Looking at $\widehat{\mathcal{F}}_{\mu\mu}(\mu)$ more closely shows

$$\begin{aligned} &\widehat{\mathcal{F}}_{\mu\mu}(\mu) \leq 0 \quad \text{for } |\mu| \leq \sqrt{\tfrac{1}{3}(\theta - \tfrac{1}{4b})}, \\ &\widehat{\mathcal{F}}_{\mu\mu}(\mu) > 0 \quad \text{for } |\mu| > \sqrt{\tfrac{1}{3}(\theta - \tfrac{1}{4b})}. \end{aligned} \tag{A.88}$$

Therefore there exists one maxima inside the ball of radius $\rho = \sqrt{\frac{1}{3}(\theta - \frac{1}{4b})}$ and two minima outside the ball of radius ρ.

Bibliography

M. Abramowitz & I. A. Stegun (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edn.

F. J. Alexander, et al. (2005). 'Accelerated Monte Carlo for Optimal Estimation of Time Series'. *Journal of Statistical Physics* **119**:1331–1345.

M. A. Alvarez, et al. (2011). 'Kernels for Vector-Valued Functions: a Review' .

B. D. O. Anderson & J. B. Moore (1979). *Optimal filtering*. Prentice-Hall, Englewood Cliffs, N.J. :.

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

C. Archambeau, et al. (2007a). 'Gaussian process approximations of stochastic differential equations'. In *Journal of Machine Learning Research Workshop and Conference Proceedings*.

C. Archambeau, et al. (2009). 'PAC-Bayes Analysis of Bayesian Inference'.

C. Archambeau & M. Opper (2011). *Approximate inference for continuous-time Markov processes*. Cambridge University Press.

C. Archambeau, et al. (2007b). 'Variational Inference for Diffusion Processes'. In *NIPS 20*, Cambridge, MA. MIT Press.

A. Argyriou, et al. (2005). 'Learning Convex Combinations of Continuously Parameterized Basic Kernels.'. In *COLT'05*, pp. 338–352.

N. Aronszajn (1950). 'Theory of reproducing kernels'. *Transactions of the American Mathematical Society* **68**.

A. Azzalini & A. Capitanio (1999). 'Statistical applications of the multivariate skew normal distribution'. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* **61**:579–602.

A. Azzalini & D. A. Valle (1996). 'The multivariate skew-normal distribution'. *Biometrika* **83**:715–726.

F. Berefelt, et al. (2003). 'Geometric Aspects of Nonlinear Filtering'. Tech. rep., Swedish Defence Research Agency.

A. Bertinet & T. C. Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers.

A. Beskos, et al. (2011). 'Hybrid Monte-Carlo on Hilbert Spaces'. *Stochastic Processes and Applications* **121**:2201–2230.

C. M. Bishop (1999). 'Latent variable models'. *Learning in Graphical Models* .

M. Briers, et al. (2004). 'Smoothing algorithms for state-space models'. Tech. rep., in Submission IEEE Transactions on Signal Processing.

M. Briers, et al. (2010). 'Smoothing algorithms for statespace models'. *Annals of the Institute of Statistical Mathematics* **62**:61–89. 10.1007/s10463-009-0236-2.

D. Brigo (2000). 'On SDEs with marginal laws evolving in finite-dimensional exponential families'. *Statistics Probability Letters* **49**(2):127–134.

D. Brigo (2011). 'Private communication'.

D. Brigo, et al. (1995). 'A Differential Geometric Approach to Nonlinear Filtering: The Projection Filter'. In *IEEE Transactions on Automatic Control*, pp. 4006–4011.

D. Brigo, et al. (1999). 'Approximate Nonlinear Filtering by Projection on Exponential Manifolds of Densities'. *Bernoulli* **5**(3):pp. 495–534.

D. Brigo, et al. (2003). 'The General Mixture Diffusion SDE and its Relationship with an Uncertain-volatility Option Model with Volatility-asset Decorrelation'. *Ssrn Electronic Journal* .

S. Canu, et al. (2002). 'Functional Learning Through Kernel'. In *Advances in Learning Theory: Methods, Models and Applications NATO Science Series III: Computer and Systems Sciences*, pp. 89–110. IOS Press.

T. K. Caughey & F. Ma (1982). 'The Steady-State Response of a Class of Dynamical Systems to Stochastic Excitation'. *Journal of Applied Mechanics* **104**(3):629–632.

W. Cheney (2001). *Analysis for Applied Mathematics*. Springer.

J. J. Collins & I. N. Stewart (1993). 'Coupled nonlinear oscillators and the symmetries of animal gaits'. *Journal of NonLinear Science* **3**:349–392.

A. P. Dempster, et al. (1977). 'Maximum Likelihood from Incomplete Data via the EM Algorithm'. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1):1–38.

A. Doucet, et al. (eds.) (2001). *Sequential Monte Carlo methods in practice*.

R. Dudley (1993). *Real Analysis and Probability*. Chapman and Hall.

G. L. Eyink (2000). 'A Variational Formulation of Optimal Nonlinear Estimation'.

G. L. Eyink, et al. (2004). 'A mean field approximation in data assimilation for nonlinear dynamics'. *Physica D: Nonlinear Phenomena* **195**(3-4):347–368.

J. Feng (2004). Chapman and Hall/CRC Press.

R. P. Feynman & A. R. Hibbs (1965). *Quantum Mechanics and Path Integrals*. McGraw-Hill Companies.

G. Folland (1984). *Real analysis*. Wiley.

S. V. Fomin & I. M. Gelfand (1963). *Calculus of Variations*. Dover Publications.

K. Friston (2008a). 'Hierarchical models in the brain'. *PLoS Computational Biology* p. 18989391.

K. Friston (2008b). 'Variational filtering'. *NeuroImage* **41**(3):747–766.

K. Friston, et al. (2006). 'A free energy principle for the brain'. *Journal of Physiology-Paris* **100**(1-3):70–87. Theoretical and Computational Neuroscience: Understanding Brain Functions.

K. Friston, et al. (2008). 'DEM: A variational treatment of dynamic systems'. *NeuroImage* **41**(3):849–885.

K. J. Friston, et al. (2010). 'Generalised Filtering'. *Mathematical Problems in Engineering* .

A. T. Fuller (1969). 'Analysis of nonlinear stochastic systems by means of the Fokker Planck equation'. *International Journal of Control* **9**(6):603–655.

Z. Ghahramani (2004). 'Unsupervised learning'. In *Advanced Lectures on Machine Learning*, pp. 72–112. Springer-Verlag.

Z. Ghahramani & S. T. Roweis (1999). 'Learning Nonlinear Dynamical Systems using an EM Algorithm'. In *Advances in Neural Information Processing Systems 11*, pp. 599–605. MIT Press.

N. J. Gordon, et al. (1993). 'Novel approach to nonlinear/non-Gaussian Bayesian state estimation'. *Radar and Signal Processing, IEE Proceedings F* **140**(2):107–113.

S. Grunewalder, et al. (2010). ‘Continuous Control with Function Indexed Gaussian Bandits’.

J. Hartikainen & S. Särkkä (2010). ‘Kalman filtering and smoothing solutions to temporal Gaussian process regression models’ pp. 379–384.

M. Havlicek, et al. (2011). ‘Dynamic modeling of neuronal responses in fMRI using cubature Kalman filtering’. *NeuroImage* .

T. Heskes & O. Zoeter (2002). ‘Expectation propagation for approximate inference in dynamic Bayesian networks’. In *In Proceedings UAI*, pp. 216–223.

M. Higgs & J. Shawe-Taylor. (2010). ‘Multiple Kernel Learning for SVM based System Identification’. In *NIPS 2010 Workshop: New Directions in Multiple Kernel Learning.*

M. Higgs & J. Shawe-Taylor (2010). ‘A PAC-Bayes Bound for Tailored Density Estimation’. In M. Hutter, F. Stephan, V. Vovk, & T. Zeugmann (eds.), *ALT*, vol. 6331 of *Lecture Notes in Computer Science*, pp. 148–162. Springer.

K. Ito (2000). ‘Gaussian Filter for Nonlinear Filtering Problems’.

T. S. Jaakkola (2000). ‘Tutorial on Variational Approximation Methods’. In *In Advanced Mean Field Methods: Theory and Practice*, pp. 129–159. MIT Press.

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

M. I. Jordan, et al. (1999a). ‘An Introduction to Variational Methods for Graphical Models’. *Mach. Learn.* **37**:183–233.

M. I. Jordan, et al. (1999b). ‘An Introduction to Variational Methods for Graphical Models’. *Mach. Learn.* **37**:183–233.

H. R. Joshi (2002). ‘Optimal control of an HIV immunology model’. *Optimal Control Appl. Methods* **23**:199–213.

S. J. Julier (1998). ‘Skewed approach to filtering’. In O. E. Drummond (ed.), *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 3373 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pp. 271–282.

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

Y. Kabanov, et al. (2006). *From Stochastic Calculus to Mathematical Finance: The Shiryaev Festschrift*. Springer.

T. Kailath (1971). ‘RKHS Approach to Detection and Estimation Problems-Part I: Deterministic Signals in Gaussian Noise’. *IEEE Transactions on Information Theory* **17**(5).

R. E. Kalman (1960). 'A New Approach to Linear Filtering and Prediction Problems'. *Journal of Basic Engineering* .

B. Kappen, et al. (2009). 'Optimal control as a graphical model inference problem' .

H. Kappen (2011). *Optimal control theory and the linear bellman equation*. Cambridge University Press.

H. J. Kappen (2005a). 'A linear theory for control of non-linear stochastic systems'. *Physical Review Letters* **95**(20):200201+.

H. J. Kappen (2005b). 'Path integrals and symmetry breaking for optimal control theory'.

P. E. Kloeden & E. Platen (1992). *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin.

A. Kolmogorov (1931). 'On Analytical Methods in the Theory of Probability'.

A. N. Kolmogorov & S. V. Fomin (1975). *Introductory Real Analysis*. Dover Publications.

F. R. Kschischang, et al. (2001). 'Factor graphs and the sum-product algorithm'. *IEEE Transactions on Information Theory* **47**:498–519.

H.-H. Kuo (2006). *Introduction to Stochastic Integration*. Springer.

H. Kushner (1967). 'Approximations to optimal nonlinear filters'. *IEEE Transactions on Automatic Control* **12**:546–556.

H. J. Kushner (2008). 'Numerical Approximations to Optimal Nonlinear Filters' .

E. S. M. L. M. Castro & R. B. Arellano-Valle (2008). 'A Note on the Identification of Skewed-Normal Distributions'.

C. Leondes, et al. (1970). 'Nonlinear Smoothing Theory'. *Systems Science and Cybernetics, IEEE Transactions on* **6**(1):63 –71.

D. Liberzon & R. W. Brockett (2000). 'Nonlinear feedback systems perturbed by noise: steady-state probability distributions and optimal control'. *IEEE Trans. Automat. Control* **45**:45–1116.

P. S. Maybeck (1979). *Stochastic models, estimation, and control*, vol. 141 of *Mathematics in Science and Engineering*.

G. Mclachlan & D. Peel (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, 1 edn.

T. Mensink, et al. (2010). 'EP for Efficient Stochastic Control with Obstacles'. In *Proceeding of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pp. 675–680, Amsterdam, The Netherlands, The Netherlands. IOS Press.

R. N. Miller, et al. (1999). 'Data assimilation into nonlinear stochastic models'.

T. P. Minka (1999). 'From Hidden Markov Models to Linear Dynamical Systems'. Tech. rep., Tech. Rep. 531, Vision and Modeling Group of Media Lab, MIT.

T. P. Minka (2000). 'Beyond Newton's method'.

T. P. Minka (2001). 'Expectation propagation for approximate Bayesian inference'. In *Uncertainty in artificial intelligence: proceedings of the seventeenth conference (2001), August 2-5, 2001, University of Washington, Seattle, Washington*, p. 362. Morgan Kaufmann.

L. Murray & A. J. Storkey (2011). 'Particle Smoothing in Continuous Time: A Fast Approach via Density Estimation'. *IEEE Transactions on Signal Processing* **59**(3):1017–1026.

M. F. Mller (1993). 'A scaled conjugate gradient algorithm for fast supervised learning'. *NEURAL NETWORKS* **6**(4):525–533.

I. T. Nabney (2002). *NETLAB: algorithms for pattern recognition*. Springer-Verlag New York, Inc., New York, NY, USA.

P. Naveau, et al. (2005). 'A skewed Kalman filter'. *J. Multivar. Anal.* **94**:382–400.

R. Neal & G. E. Hinton (1998). 'A View Of The Em Algorithm That Justifies Incremental, Sparse, And Other Variants'. In *Learning in Graphical Models*, pp. 355–368. Kluwer Academic Publishers.

R. M. Neal (2010). 'MCMC Using Hamiltonian Dynamics'. In G. L. J. Steve Brooks, Andrew Gelman & X.-L. Meng (eds.), *Handbook of Markov Chain Monte Carlo*, chap. 5. Chapman & Hall/CRC.

B. Øksendal (2003). *Stochastic Differential Equations: An Introduction with Applications (Universitext)*. Springer, 6th edn.

M. Opper (1998). *A Bayesian approach to on-line learning*, pp. 363–378. Cambridge University Press, New York, NY, USA.

J. Ormerod (2011). 'Skew-Normal Variational Approximations for Bayesian Inference'. Tech. rep., School of Mathematics and Statistics, University of Sydney.

T. Ozaki (1992). 'A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach'. *Statistica Sinica* **2**(1):113–135.

T. Ozaki (1993). 'A local linearization approach to nonlinear filtering'. *International Journal of Control* **57**(1):75–96.

E. Parzen (1961). 'An Approach to Time Series Analysis'. *The Annals of Mathematical Statistics* **32**.

M. K. Pitt & N. Shephard (1999). 'Filtering via Simulation: Auxiliary Particle Filters'. *Journal of the American Statistical Association* **94**(446):590–599.

G. d. Prato (2006). *An Introduction to Infinite-Dimensional Analysis*. Springer.

G. D. Prato & J. Zabczyk (1992). *Stochastic Equations in Infinite Dimensions*, vol. 44. Cambridge University Press, In Encyclopedia of Mathematics and Its Applications.

C. E. Rasmussen & C. K. I. Williams (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

H. E. Rauch, et al. (1965). 'Maximum Likelihood Estimates of Linear Dynamic Systems'. *Journal of the American Institute of Aeronautics and Astronautics* **3**:1445–1450.

J. Restrepo (2008). 'A path integral method for data assimilation'. *Physica D: Nonlinear Phenomena* **237**(1):14–27.

J. J. Riera, et al. (2004). 'A state-space model of the hemodynamic approach: nonlinear filtering of BOLD signals'. *Neuroimage* pp. 547–567.

H. Risken (1996). *The Fokker-Planck Equation: Methods of Solutions and Applications*. Springer Series in Synergetics. Springer, 2nd ed. 1989. 3rd printing edn.

J. Roberts & P. Spanos (2003). *Random Vibration and Statistical Linearization*. Dover Publications.

S. Särkkä (2007). 'On Unscented Kalman Filtering for State Estimation of Continuous-Time Nonlinear Systems'. *IEEE Transactions on Automatic Control* **52**(9):1631–1641.

S. Särkkä (2010). 'Continuous-time and continuous-discrete-time unscented Rauch-Tung-Striebel smoothers'. *Signal Process.* **90**:225–235.

S. Särkkä & J. Hartikainen (2010a). 'On Gaussian Optimal Smoothing of Non-Linear State Space Models'. *IEEE Transactions on Automatic Control* **55**:1938–1941.

S. Särkkä & J. Hartikainen (2010b). 'Sigma Point Methods in Optimal Smoothing of Non-Linear Stochastic State Space Models'. *IEEE International Workshop on Machine Learning for Signal Processing* .

S. Särkkä & J. Sarmavouri (2011). 'Gaussian Filtering and Smoothing for Continuous-Discrete Dynamic Systems'. *Under review* .

S. Schaal (2006). 'Dynamic Movement Primitives -A Framework for Motor Control in Humans and Humanoid Robotics'. pp. 261–280.

B. Scholkopf & A. J. Smola (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

M. Schumaker (1987). 'Center manifold reduction and normal form transformations in systems with additive noise'. *Physics Letters A* **122**(6-7):317 – 322.

Y. Shen, et al. (2010). 'A Comparison of Variational and Markov Chain Monte Carlo Methods for Inference in Partially Observed Stochastic Dynamic Systems'. *Journal of Signal Processing Systems* **61**:51–59. 10.1007/s11265-008-0299-y.

J. R. Shewchuk (1994). 'An introduction to the conjugate gradient method without the agonizing pain'. Tech. rep., School of computer science, Carnegie Mellon University.

A. J. Smola, et al. (1998). 'The connection between regularization operators and support vector kernels'. *Neural Netw.* **11**(4):637–649.

Y. Steinberg, et al. (1988). 'On the optimal filtering problem for the cubic sensor'. *Circuits, Systems, and Signal Processing* **7**:381–408. 10.1007/BF01599977.

F. Steinke & B. Schölkopf (2008). 'Kernels, regularization and differential equations'. *Pattern Recogn.* **41**(11):3271–3286.

F. Steinke & B. Schlkopf (2006). 'MACHINE LEARNING METHODS FOR ESTIMATING OPERATOR EQUATIONS'. *14th IFAC Symposium on System Identification* .

O. Stramer & R. L. Tweedie (1999). 'Langevin-Type Models I: Diffusions with Given Stationary Distributions and their Discretizations*'. *Methodology and Computing in Applied Probability* **1**(3):283–306.

C. T. Striebel (1965). 'Partial differential equations for the conditional distribution of a Markov process given noisy observations'. *Journal of Mathematical Analysis and Applications* **11**:151 – 159.

A. Stuart (2010). 'Inverse problems: a Bayesian perspective'. In *Acta Numerica 2010*, vol. 17.

E. Theodorou, et al. (2010). 'a generalized path integral control approach to reinforcement learning' (11):3137–3181.

L. Tierney & J. B. Kadane (1986). 'Accurate Approximations for Posterior Moments and Marginal Densities'. *Journal of the American Statistical Association* **81**(393):82–86.

E. Todorov (2008). 'General duality between optimal control and estimation'. In *CDC*, pp. 4286–4292. IEEE.

A. W. van der Vaart & J. H. van Zanten (2008). *Reproducing kernel Hilbert spaces of Gaussian priors*, vol. 3 of *IMS Collections*, pp. 200–222.

N. G. van Kampen (2007). *Stochastic Processes in Physics and Chemistry*. North-Holland, 3 edn.

P. J. van Leeuwen (2010). 'Nonlinear data assimilation in geosciences: an extremely efficient particle filter'. *Q.J.R. Meteorol. Soc.* **136**(653):1991–1999.

M. D. Vrettas (2010). *Approximate Bayesian techniques for inference in stochastic dynamical systems*. Ph.D. thesis, Aston University.

M. D. Vrettas, et al. (2010). 'A new variational radial basis function approximation for inference in multivariate diffusions'. *Neurocomput.* **73**:1186–1198.

M. J. Wainwright & M. I. Jordan (2008). 'Graphical Models, Exponential Families, and Variational Inference'. *Found. Trends Mach. Learn.* **1**:1–305.

J. Weickert (2001). 'Applications of nonlinear diffusion in image processing and computer vision'. *Acta Math. Univ. Comenianae* **70**:33–50.

D. J. Wilkinson (2006). *Stochastic Modelling for Systems Biology (Chapman & Hall/CRC Mathematical & Computational Biology)*. Chapman and Hall/CRC, 1 edn.

B. M. Yu, et al. (2007). 'Neural decoding of movements: From linear to nonlinear trajectory models'. In M. Ishikawa, K. Doya, H. Miyamoto, & T. Yamakawa (eds.), *Neural Information Processing (ICONIP 2007), Part I*, vol. 4984, pp. 586–595. Springer-Verlag Berlin Heidelberg. ISBN 978-3-540-69154-9.

B. M. Yu, et al. (2006). 'M.: Expectation propagation for inference in non-linear dynamical models with Poisson observations'. In *In: Proc IEEE Nonlinear Statistical Signal Processing Workshop. (2006*.

Y. Zhao, et al. (2006). 'General Design Bayesian Generalized Linear Mixed Models'. *Statistical Science* **21**(1):35–51.

O. Zoeter & T. Heskes (2005). 'Gaussian Quadrature Based Expectation Propagation'. In R. Cowell & Z. Ghahramani (eds.), *International Workshop on Artificial Intelligence and Statistics (AISTATS)*, vol. 10.