Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Prakhar Verma

Sparse Gaussian processes for stochastic differential equations

Master's Thesis
Espoo, July 30, 2021

Supervisor: Assistant Professor Arno Solin, Aalto University
Advisor: Dr. Vincent Adam, Aalto University

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author: Prakhar Verma

Title:
Sparse Gaussian processes for stochastic differential equations

Date:	July 30, 2021	**Pages:**	60
Major:	Machine Learning, Data Science, and Artificial Intelligence	**Code:**	SCI3044
Supervisor:	Assistant Professor Arno Solin		
Advisor:	Dr. Vincent Adam, Aalto University		

Dynamical systems present in the real world are often well represented using stochastic differential equations (SDEs) incorporating the sources of stochasticity. With the recent advances in machine learning (ML), research has been done to develop algorithms to learn SDEs based on observations of dynamical systems.

The thesis frames the SDE learning problem as an inference problem and aims to maximize the marginal likelihood of the observations in a joint model of the unobserved paths and the observations through an observation model. As this problem is intractable, a variational approximate inference algorithm is employed to maximize a lower bound to the log marginal likelihood instead of the original objective. In the variational framework, Gaussian processes (GPs) have been used as approximate posterior over paths. However, the resulting algorithms require fine discretization of the time horizon resulting in high complexity.

The recent advances related to exploiting sparse structure in the GPs are explored in the thesis, and an alternate parameterization of the approximate distribution over paths using a sparse Markovian Gaussian process is proposed. The proposed method is efficient in storage and computation, allowing the usage of well-established optimizing algorithms such as natural gradient descent. The capability of the proposed method to learn the SDE from observations is showcased in the two experiments: the Ornstein–Uhlenbeck (OU) process and a double-well process.

Keywords: stochastic differential equations, sparse Gaussian processes, variational inference, natural gradient descent, dynamic systems

Language: English

ज्ञानं परमं बलम्||

- *Knowledge is the supreme power*

Acknowledgements

Studying abroad requires a lot of support, and I will forever be grateful to my family and friends back in India without the support of whom this would not have been possible.

Thanks to my previous colleagues at TomTom, who helped me realize my interest in machine learning and motivated me to pursue academics in it.

I am deeply indebted to Professor Arno Solin, with whom the past year has been full of learning. From being meticulous about things to being patient in research, I have learnt a lot. Thanks to Vincent Adam for introducing me to the topic and guiding me throughout this work. There were meetings when I felt lost, and it is he who put me on the right track. Thanks to both for making this thesis memorable!

The work was done in the AaltoML group, and I would like to thank every group member. A special shout-out to Paul Chang for providing feedback on the work and William Wilkinson for helping with the implementation.

Last but definitely not least, thanks to all the people at Aalto with whom I had a conversation during my master's. I have learnt something from each one of you!!!

Espoo, July 30, 2021

Prakhar Verma

Abbreviations and Acronyms

ELBO evidence lower bound

GP Gaussian process

GPR Gaussian process regression

GP-SDE Gaussian process variational approximation for stochastic differential equation (Archambeau et al., 2007, 2008)

IID independent and identically distributed

IVP initial value problem

KL Kullback–Leibler divergence

LTI linear time invariant

ML machine learning

ODE ordinary differential equation

OU Ornstein–Uhlenbeck process

S^{2}VGP doubly sparse variational Gaussian process

SDE stochastic differential equation

SGD stochastic gradient descent

SGP-SDE Sparse Gaussian process stochastic differential equation, the proposed method

SSM state-space model

SVGP sparse variational Gaussian process

VI variational inference

Contents

Chapter 1

Introduction

A dynamical system is a system whose configuration or state evolves following a differential equation. It is present in abundance in the real world, from population growth and celestial body to healthcare (Eraker, 2001; Golightly & Wilkinson, 2011; van Kampen, 2007). If the system is deterministic, it may be described by an ordinary differential equation (ODE). However, most real-world dynamical systems are stochastic, best described by a stochastic differential equation (SDE), further discussed in Section 2.3.

Learning about a dynamical system gives insights about it, helping optimize the system, evaluate it, and even in decision-making. However, it is a challenging task. A dynamical system may have two sources of stochasticity; one from the describing SDE and another from the observation model. Both of these sources can be complex, leading to intractability. Often, the observations of a dynamical system are at discrete time intervals, and the aim is to learn a continuous system adding to the complexity of the task.

Machine learning (ML), the science of algorithms that learn from data over time, is ubiquitous nowadays, with applications ranging from healthcare, finance to education impacting millions of people across the globe (Bhardwaj et al., 2017; Heaton et al., 2018; Ciolacu et al., 2017). With the recent advances in ML, research has been done to develop new algorithms to learn the SDE based on the observations of a dynamical system. Bayesian methods have been quite popular for stochastic systems due to their inherent property of handling stochasticity. Gaussian processes (GPs) (Rasmussen & Williams, 2006) are used extensively for dynamical systems due to their well-established connection with SDEs (Särkkä & Solin, 2019, Chapter 12).

A stochastic differential equation (SDE) consists of a deterministic term, drift, and a stochastic term, diffusion. A dynamical system with linear SDEs and a Gaussian observation model leads to a closed-form expression for the posterior for the state trajectory using Bayesian methods and conjugate prop-

erties. However, for non-Gaussian likelihoods, approximate inference algorithms are employed to approximate the intractable posterior. Non-linear SDEs involve more non-conjugate terms leading to sophisticated approximate algorithms to make the problem tractable. Variational inference (VI) is one of the popular approximate inference methods in statistical machine learning, discussed in Section 2.2. Many state-of-the-art methods to learn SDEs employ VI, which involves optimizing an evidence lower bound (ELBO). An essential step in these methods is the fine discretization of the time horizon, leading to a high number of parameters.

The thesis aims to develop a method to learn the SDE describing a dynamical system with an arbitrary likelihood based on a set of noisy observations. The problem is framed as an inference problem, and variational inference (VI) is used. A novel alternate parameterization to the approximate distribution over paths is proposed using a sparse Markovian Gaussian process, inspired by the doubly sparse Gaussian process (Adam et al., 2020), discussed in Section 2.6. It reduces the complexity of the method both in storage and time, allowing the usage of well-established optimizing algorithms such as natural gradient descent. As SDEs are continuous over time, ELBO differs from that of the standard VIs prominently used in machine learning tasks and is derived incorporating the continuous nature of SDEs.

The thesis is structured as follows: Chapter 2 discusses the generative models, SDEs, GPs, approximating methods, and sets a base for the following chapters. In Chapter 3, the recent research work is discussed, and an in-depth review of two closely related research papers is done. Chapter 4 introduces the proposed method, model specifications, the objective for variational inference (VI), ELBO, is derived, and the natural gradient-based optimization algorithm is presented. Chapter 5 discusses the experiments on two processes: the Ornstein–Uhlenbeck (OU) and a double-well process. The posterior obtained for the two processes are evaluated and compared with other models in Chapter 6, and plausible future works are discussed. Chapter 7 rounds off with a conclusion.

Chapter 2

Background

This chapter discusses the fundamental concepts that will set the stage for the further chapters discussing different methods and models with the primary aim of learning and performing inference in a dynamic system with SDE priors. The chapter starts by introducing generative models in Section 2.1 and inference in these models. Then, approximate inference as an inference method is discussed to deal with intractability, and Kullback–Leibler (KL) divergence and variational inference (VI) are introduced in Section 2.2. In Section 2.3, dynamic systems are introduced, followed by a brief introduction to SDEs, the use of SDE priors in learning the dynamic systems, SSMs, and Markovian Gaussian process. The chapter concludes with a discussion on complexity in Gaussian processes reviewing SVGP in Section 2.5 and S^{2}VGP in Section 2.6.

2.1 Generative models

Formally, machine learning models are divided into two types: generative models and discriminative models. A generative model is one where the joint distribution of the observed variable and the target variable is learnt however in discriminative models the conditional distribution between them is learnt (Ng & Jordan, 2002).

Suppose, $\mathbf{X}$ is our observed variable and $\mathbf{Y}$ is the target variable. Then,

1. **Generative model** focuses on learning $p(\mathbf{X}, \mathbf{Y})$
2. **Discriminative model** focuses on learning $p(\mathbf{Y} \mid \mathbf{X})$

From the definition of both the models, it can be inferred that the generative models are used to generate more data however discriminative models is used

only to get the target value. Alternatively, generative models are also defined as the ones which aim to learn $p(\mathbf{X} \mid \mathbf{Y})$.

Using Bayes' rule, generative model is written as

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

where $p(\mathbf{X} \mid \mathbf{Y})$ is the posterior distribution, $p(\mathbf{Y} \mid \mathbf{X})$ is the likelihood distribution, $p(\mathbf{X})$ is the prior distribution, and $p(\mathbf{Y})$ is the marginal likelihood. As $p(\mathbf{Y})$ is a constant, the posterior distribution is written as

$$p(\mathbf{X} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \mathbf{X})\, p(\mathbf{X}). \tag{2.2}$$

In generative models, there are two ways to perform inference: exact and approximate inference. Exact inference is possible only under certain conditions; when the unobserved state is limited and discrete, or conjugate pairs are used which makes the calculation of posterior possible in closed form. However, most of the time exact inference is not tractable so approximate inference is used (Bishop, 2006, Chapter 10).

2.2 Approximate inference

Approximate inference is one of the inference methods where the exact posterior distribution is approximated rather than solving it in exact form. The two most popular methods for approximate inference are Monte Carlo methods and variational inference (VI). Monte Carlo method is a sampling-based method, samples are drawn from the posterior distribution to approximate it, however in variational inference the posterior distribution is approximated with another parametric distribution (Bishop, 2006, Chapter 10, 11).

In this thesis, the focus is on VI and thus it is discussed in detail. However, before discussing VI, KL divergence is introduced as it is an important component of it.

2.2.1 Kullback–Leibler (KL) divergence

KL divergence is used to measure distance between two distributions. Mathematically, it is defined as

$$\mathrm{D}_{\mathrm{KL}}\left[P \,\|\, Q\right] = \int q(\mathbf{x}) \log \tfrac{q(\mathbf{x})}{p(\mathbf{x})}\, \mathrm{d}\mathbf{x}, \tag{2.3}$$

where, $p(\mathbf{x})$ and $q(\mathbf{x})$ are two probability density functions of the two distributions P and Q.

From Eq. (2.3), it is inferred that KL divergence is asymmetric and does not follow triangle inequality. Also, $\mathrm{D_{KL}} >= 0$ and is zero only when $p(\mathbf{x}) == q(\mathbf{x})$.

Example: Suppose, the true distribution, P, is a Gaussian distribution with mean 0 and variance 0.5. Three distributions, namely Q_1, Q_2, Q_3, with mean $0.5, 0.2, 0$ and variance $0.35, 0.5, 0.45$ respectively are available and the goal is to find the distribution which is closest to the true distribution P.

Figure 2.1 showcases the three probable distributions along with the KL divergence values and the true distribution. It can be inferred both from the plots and from the KL divergence value that Q_3 is the closest to P.

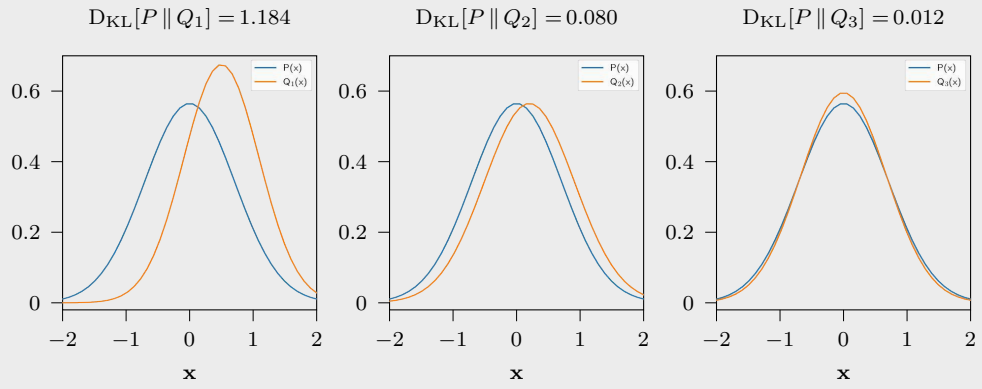

Figure 2.1: True distribution, P, along with the probable distributions, $\mathrm{Q_1}$, $\mathrm{Q_2}$, $\mathrm{Q_3}$, and their KL divergence values.

2.2.2 Variational inference

Variational inference (VI) is a method of approximate inference where an intractable distribution is approximated with a tractable distribution.

Suppose, P is an intractable distribution and is approximated by a tractable distribution, Q, belonging to a particular family of distributions. To compare the two distributions, KL divergence is used. Thus, the aim is to minimize the KL divergence between these two distributions.

Example: Suppose, there is a model that is projecting the data from the observation space to the latent space. The aim is to learn a distribution, $p(\mathbf{Z} \mid \mathbf{Y})$, where $\mathbf{Y}$ is the data in the observation space and $\mathbf{Z}$ is the corresponding data in the latent space. Using Bayes' theorem, the interested distribution is written as

$$p(\mathbf{Z} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathbf{Z})\, p(\mathbf{Z})}{p(\mathbf{Y})}, \tag{2.4}$$

where $p(\mathbf{Y} \mid \mathbf{Z})$ is the likelihood, $p(\mathbf{Z})$ is the prior, and $p(\mathbf{Y})$ is the marginalized likelihood, which is a constant.

The goal is to approximate the posterior distribution with $q(\mathbf{Z})$. To compare the two distributions, true and approximating, KL divergence is used with the objective to minimize it. Using Eq. (2.3) and Eq. (2.4), the KL is written as

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{Z}) \,\|\, p(\mathbf{Z} \mid y)\right] &= \mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z}) - \log p(\mathbf{Z} \mid \mathbf{Y})] \\
&= \mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z}) - \log p(\mathbf{Y} \mid \mathbf{Z}) - \log p(\mathbf{Z}) + \log p(\mathbf{Y})] \\
&= \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{Z}) \,\|\, p(\mathbf{Z})\right] - \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{Y} \mid \mathbf{Z})] + \log p(\mathbf{Y}) \\
\log p(\mathbf{Y}) &= -\mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{Z}) \,\|\, p(\mathbf{Z})\right] + \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{Y} \mid \mathbf{Z})] \\
&\quad + \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{Z}) \,\|\, p(\mathbf{Z} \mid \mathbf{Y})\right]
\end{aligned} \tag{2.5}$$

As $\log p(\mathbf{Y})$ is a constant, from Eq. (2.5), it is inferred that minimizing $\mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{Z}) \,\|\, p(\mathbf{Z} \mid \mathbf{Y})\right]$ is equivalent to maximizing the sum of the other two terms, $\mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{Y} \mid \mathbf{Z})] - \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{Z}) \,\|\, p(\mathbf{Z})\right]$, also known as evidence lower bound (ELBO).

Thus, the objective of minimizing the KL can also be framed as maximizing the ELBO.

2.3 Dynamic system

Dynamic system is a system where the motion occurs; components or variables evolve over time. Suppose, there is a dynamic system where a variable $\mathbf{y}$ is observed over time t. Further, there is a hidden state $\mathbf{x}$ that is unobserved which governs the evolution of $\mathbf{y}$ over time. Thus, using Bayes' theorem, the model is written as

$$p(\mathbf{x}_t \mid \mathbf{y}_t) \propto p(\mathbf{y}_t \mid \mathbf{x}_t)\, p(\mathbf{x}_t),$$

where $\mathbf{y}$ is observed through an observation model. If Gaussian, the observation model is written as

$$p(\mathbf{y}_t \mid \mathbf{x}_t) \sim \mathcal{N}(\mathbf{y}_t \mid \mathbf{H}\,\mathbf{x}_t,\, \mathbf{R}),$$

where $t \in [0,\, T]$(McGoff et al., 2015).

Furthermore, for inference, a prior needs to be introduced over $\mathbf{x}$. For dynamic modelling, differential equations become a natural option for priors. Due to the robustness of SDEs over ODEs, SDEs are more preferred. However, before discussing the models with SDE priors, SDEs are briefly discussed.

2.3.1 Stochastic differential equation (SDE)

Differential equations are the equations relating functions with their derivatives with applications ranging in fields including physics, chemistry, and economics. The dynamics of any system can be expressed in the form of a differential equation governing the change of the state of the system over time:

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathrm{f}(\mathbf{x}(t),\, t), \tag{2.6}$$

where $\mathbf{x}(t)$ is the state of the system at time t and f(.) is the governing function. This is known as an *ordinary differential equation (ODE)* (Griffiths & Higham, 2010).

Given the initial value, Eq. (2.6) is solved by:

$$\mathbf{x}(T) = \mathbf{x}(0) + \textstyle\int_{t=0}^{T} \mathrm{f}(\mathbf{x}(t),\, t)\,\mathrm{d}t, \tag{2.7}$$

which is commonly known as an initial value problem (IVP). A numerical method used to solve IVP is the *Euler's Method*

$$\mathbf{x}_{t+h} = \mathbf{x}_t + \mathrm{f}(\mathbf{x}(t),\, t)\,h, \tag{2.8}$$

where h is the time-step(Griffiths & Higham, 2010, Chapter 2).

Example: Consider a spring-mass model

$$\frac{\mathrm{d}^2\mathbf{x}(t)}{\mathrm{d}t^2} + \gamma\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} + v^2\mathbf{x}(t) = w(t). \tag{2.9}$$

In the state-space form, it is written as:

$$\mathrm{f}(\mathbf{X}(t), t) = \begin{pmatrix} 0 & 1 \\ -v^2 & -\gamma \end{pmatrix} \mathbf{X}(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w(t), \tag{2.10}$$

where $\mathrm{f}(\mathbf{X}(t), t) = \frac{\mathrm{d}\mathbf{X}(t)}{\mathrm{d}t}$ and $\mathbf{X}(t) = \begin{pmatrix} \mathbf{x}(t) \\ \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} \end{pmatrix}$.

Let the parameters be $v = 2, \gamma = 1$ and $w(t) = 0$. The trajectory simulated using Euler's method for $t \in [0, 10]$ with step-size 0.05 and initial position $\mathbf{x}_0 = [0, 1]$ is shown in Figure 2.2.

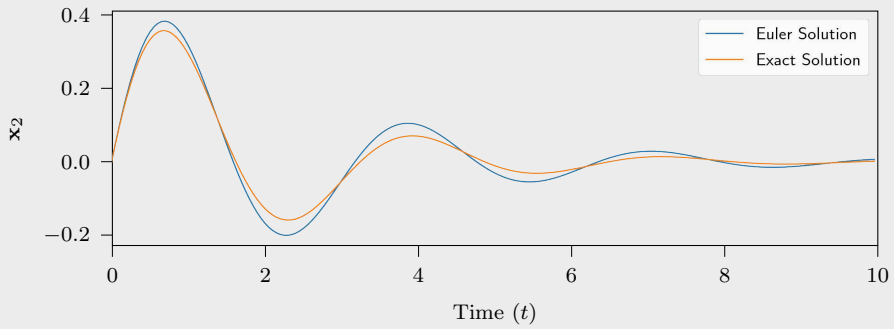

Figure 2.2: Euler solution for the spring-mass IVP model and its comparison with the true solution.

One of the drawbacks associated with ODEs is that it does not include the possible uncertainty that might be present in the system/environment/sensors. One of the plausible ways of including them is adding Gaussian noise

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathrm{F}(\mathbf{x}(t),\, t) + \boldsymbol{\epsilon}, \tag{2.11}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\, \mathbf{I})$. In Eq. (2.11), the intensity of noise is constant which can be made variable by introducing a new function $\mathrm{L}(\cdot)$

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathrm{F}(\mathbf{x}(t),\, t) + \mathrm{L}(\mathbf{x}(t), t)\, \boldsymbol{\epsilon}. \tag{2.12}$$

A problem associated with Eq. (2.12) is that it is not globally differentiable because of the discontinuous Gaussian white noise. Itô integral is solved to mitigate this issue which leads to the following equation, with the use of Brownian motion,

$$\mathrm{d}\mathbf{x}(t) = \mathrm{F}(\mathbf{x}(t), t)\, \mathrm{d}t + \mathrm{L}(\mathbf{x}(t), t)\, \mathrm{d}\boldsymbol{\beta}(t), \tag{2.13}$$

where $\mathrm{F}(\mathbf{x}(t),\, t)$ is known as a drift function, $\mathrm{L}(\mathbf{x}(t), t)$ as a diffusion function, and $\mathrm{d}\boldsymbol{\beta}(t)$ is a Brownian motion with spectral density $\mathbf{Q}$. A few details in the derivation is omitted which can be found in Särkkä & Solin (2019, Chapter 3).

A common numerical method to solve an SDE, Eq. (2.13), is *Euler–Maruyama*, (Sauer, 2012)

$$\mathbf{x}_{t+h} = \mathbf{x}_t + \mathrm{F}(\mathbf{x}_t, t)\, h + \sqrt{h}\, \mathrm{L}(\mathbf{x}_t, t)\, \boldsymbol{\omega}(t), \tag{2.14}$$

where $\boldsymbol{\omega}(t) \sim \mathcal{N}(0, \mathbf{Q})$ and h is the time-step.

Example: In Eq. (2.10), if the driving force is considered as a Gaussian noise, the equation is written as

$$\mathrm{f}(\mathbf{X}(t), t) = \begin{pmatrix} 0 & 1 \\ -v^2 & -\gamma \end{pmatrix} X(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \boldsymbol{\epsilon}(t), \tag{2.15}$$

where $\boldsymbol{\epsilon}(t) \sim \mathcal{N}(0, \mathbf{Q})$. Now, by employing Euler–Maruyama with the same values as in the previous example and $\mathbf{Q} = 0.005\,\mathbf{I}$, the trajectories are simulated as shown in Figure 2.3.

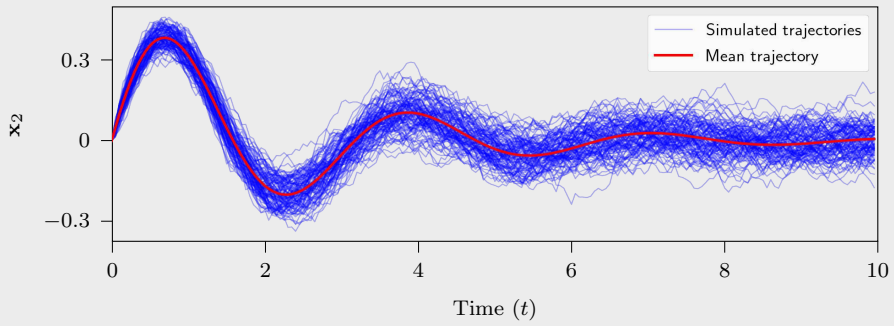

Figure 2.3: 100 Euler–Maruyama solution trajectories for the spring-mass IVP model with Gaussian noise and the mean trajectory.

SDEs give stochastic solutions(trajectories) as compared to the deterministic solution of an ODE. Thus, one of the main advantages of SDEs over ODEs is its robustness and capability to quantify uncertainty.

2.3.2 SDE priors

As discussed in Section 2.3, a dynamic system with Gaussian observation model is written as

$$p(\mathbf{x}_t \mid \mathbf{y}_t) \propto p(\mathbf{y}_t \mid \mathbf{x}_t)\, p(\mathbf{x}_t),$$
$$p(\mathbf{y}_t \mid \mathbf{x}_t) \sim \mathcal{N}(\mathbf{H}\,\mathbf{x}_t, \mathbf{R}).$$

Further, to learn the dynamics, an SDE prior over $\mathbf{x}$ is introduced

$$\mathrm{d}\mathbf{x}(t) = \mathrm{f}(\mathbf{x}(t), t)\, \mathrm{d}t + \mathrm{L}(\mathbf{x}(t), t)\, \mathrm{d}\boldsymbol{\beta}(t).$$

With SDE prior, one assumption made is that the system is Markovian. By Markovian, it is meant that at any time t the conditional probability of the future event given the entire past is same as the conditional probability of that future event given the state at time t.

To perform inference, that is calculate the posterior, approximate inference methods can be used. As discussed in Särkkä & Solin (2019), SDEs with a linear drift function have a relation with GPs. Thus, with an assumption of linear SDE and using the property of Markovian, VI can be performed with the family of approximating distribution to be Markovian GPs.

2.3.3 State-space models

State-space model (SSM) is a modelling method where the state of a dynamic system is represented as a set of first order differential or difference equations. A state of a system is defined as the minimal set of variables that fully describe the system that is it has enough information to predict the future state/behaviour (Solin, 2016, Chapter 3).

Example: Consider a linear time-variant dynamic model, where $\mathbf{x} \in \mathbb{R}^n$ is the state vector, $\mathbf{y} \in \mathbb{R}^m$ is the observed vector, and velocity of $\mathbf{x}$ is represented as $\dot{\mathbf{x}}(t)$. Then, the state-space model equations are written as

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}(t)\,\mathbf{x}(t) + \mathbf{B}(t)\,\mathbf{u}(t),\\ \mathbf{y}(t) &= \mathbf{C}(t)\,\mathbf{x}(t) + \mathbf{D}(t)\,\mathbf{u}(t),\end{aligned}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the state vector, $\mathbf{y}(t) \in \mathbb{R}^m$ is the observed vector, $\mathbf{u}(t) \in \mathbb{R}^p$ is the control vector, $\mathbf{A}(t) \in \mathbb{R}^{n\times n}$ is the system matrix, $\mathbf{B}(t) \in \mathbb{R}^{n\times p}$ is the input matrix, $\mathbf{C}(t) \in \mathbb{R}^{m\times n}$ is the output matrix, $\mathbf{D}(t) \in \mathbb{R}^{m\times p}$ is the feedback matrix.

SSMs are written for both continuous and discrete time models as well as time-variant and time-invariant models. One of the prime advantages of SSM is the compact and concise representation of the system which helps in quick and efficient analysis.

2.4 Markovian Gaussian process

A Gaussian process (GP) (Rasmussen & Williams, 2006) is a distribution over functions. Formally, a GP is defined as a random function, $f(\mathbf{x})$, on

an input space $\mathbb{R}^d$ characterized by a mean function, $\mu(\mathbf{x})$, and a covariance function, $\kappa(\mathbf{x}, \mathbf{x}')$. The GP prior over $f(\mathbf{x})$ is written as

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')), \tag{2.16}$$

where $\mu(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ and $\kappa(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}(f(\mathbf{x}), f(\mathbf{x}'))$.

As discussed, a Markovian process has the distribution of $\mathbf{x}_{n+1}$ condition on $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n\}$ same as the distribution condition on only $\mathbf{x}_n$.

Thus, a stochastic process is a Markovian Gaussian process if it follows the properties of both Gaussian and a Markovian process (Rasmussen & Williams, 2006, Appendix B).

2.5 Sparse variational Gaussian process

One of the major limitations of a GP model is the computational complexity as it involves inverting a $\mathbb{R}^{n \times n}$ matrix which is an $O(n^3)$ operations. Thus, it scales poorly with data.

Lately, research has progressed in this direction with one of the most prominent work being sparse variational Gaussian process (SVGP) (Hensman et al., 2015b) which uses variational approximation. It uses m inducing points rather than the actual n data points; inducing points are not necessarily data points and are learnt over time. It reduces the computational complexity to $O(m^3)$. The augmented model with inducing points is written as $p(\mathbf{y},\, f,\, \mathbf{u}) = p(\mathbf{y} \mid f)\, p(f \mid \mathbf{u})\, p(\mathbf{u})$.

Variational inference (VI) is used to approximate the posterior $p(f,\, \mathbf{u} \mid \mathbf{y})$ by $q(f,\, \mathbf{u})$ and the family of distribution Q is chosen to be of the form $q(f,\, \mathbf{u}) = p(f \mid \mathbf{u})\; q(\mathbf{u})$ where $q(\mathbf{u}) \sim \mathcal{N}(\mathbf{m}, \mathbf{S})$.

The ELBO can be calculated using Jensen's inequality as

$$\begin{aligned} \log p(\mathbf{y}) &\geq \mathbb{E}_{q(\mathbf{u},\, f)} \left[\log p(\mathbf{y} \mid f)\right] - \mathbb{E}_{q(\mathbf{u},\, f)} \left[\log \frac{q(f,\, \mathbf{u})}{p(f,\, \mathbf{u})}\right] \\ &\geq \mathbb{E}_{q(f)} \left[\log p(\mathbf{y} \mid f)\right] - \mathrm{D}_{\mathrm{KL}} \left[q(\mathbf{u}) \,\|\, p(\mathbf{u})\right]. \end{aligned} \tag{2.17}$$

Example: There is a set of 200 data points and the goal is to condition a GP model on it. However, due to high complexity of a classical GP model, SVGP with 20 learnable inducing points is used.

The ELBO is optimized for 40 epochs using Adam optimizer (Kingma & Ba, 2015) with learning rate 0.05.

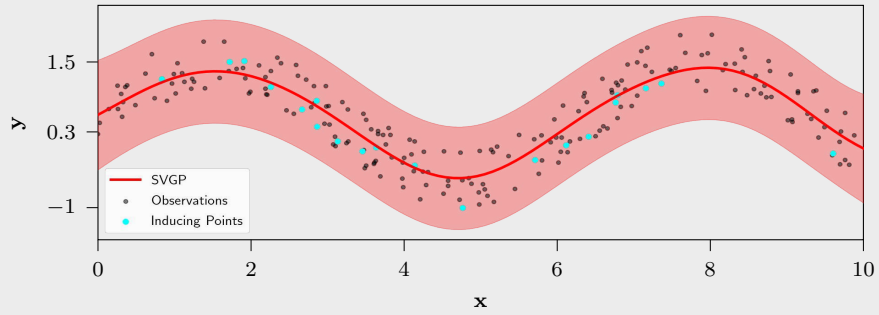

Figure 2.4: Mean and 95% confidence interval of the posterior of the SVGP model along with the learnt inducing points(cyan) and the data points(black)

2.6 Doubly sparse variational Gaussian process

Adam et al. (2020) researched on combining the benefits of SSM and SVGP to tackle the problem of computational complexity in GPs. They combine the idea of inter-domain inducing features with the state-space GP formulation referring to it as doubly sparse variational Gaussian process (S^2VGP). The linear operator, ψ, for inducing features is chosen as

$$\psi_i : f \to s(\mathbf{z}_i) = [f(\mathbf{z}_i), \ldots, f^{(d-1)}(\mathbf{z}_i)]^\top ,$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_m]$ are the *ordered* inducing points and $s(\cdot)$ represent the states.

This leads to the inducing states $\mathbf{U} = \{\psi_i[f]\}_{i=1}^M$ be Markovian and follow a Gaussian distribution that is $p_\psi(\mathbf{U}) = p(\mathbf{u}_1) \prod_{i=1}^M p(\mathbf{u}_{i+1} \mid \mathbf{u}_i)$. Another interesting property is that the posterior of the function is dependent only on the closest right and left inducing points, $p(f(\mathbf{x}_n) \mid \mathbf{u}) = p(f(\mathbf{x}_n) \mid \mathbf{u}_{n-}, \mathbf{u}_{n+})$. The ELBO of S^2VGP is defined as

$$\ell = \textstyle\sum_n \mathbb{E}_q[\log p(\mathbf{y}_n \mid f(\mathbf{x}_n))] - \frac{1}{2}\mathrm{tr}(\mathbf{Q}_\psi \, \Sigma_{\mathbf{uu}}) + \frac{1}{2}|\Sigma_{\mathbf{uu}}| + c(\mu_{\mathbf{u}}, p_\psi). \quad (2.18)$$

For more details, reader is advised to go through Adam et al. (2020).

Chapter 3

Related work

This chapter discusses the most prominent and recent research work related to the learning of an SDE. It is broadly divided into two sections: the first section discusses the related research work and the second section discusses two closely related research papers in detail.

3.1 Literature review

Gaussian processes (GPs) (Rasmussen & Williams, 2006) provide an elegant statistical machine learning framework that estimates uncertainty. However, they are infamous for their high complexity, scaling as $O(n^3)$ in time and $O(n^2)$ in space, making them impractical for numerous datasets consisting of thousands of observations/data points. However, researchers have been quite active in this area to mitigate this issue. One of the prominent works is removing redundant information and exploiting the sparse structure by introducing inducing features (Hensman et al., 2015b), reducing the complexity to $O(m^3)$, m being the number of inducing features. As $m << n$, it makes GPs more scalable, easy to train, and usable even for large datasets.

Generally, sparse GPs do not have a closed-form solution, and thus approximate inference is used (Titsias, 2009). Variational inference (VI) is one of the most popular methods where the inference problem is cast as an optimization problem. Other methods researched are Markov-chain-Monte-Carlo methods, expectation propagation (Hensman et al., 2015a; Bui et al., 2017).

Adam et al. (2020) researched on combining the benefits of SSM and sparse GPs using variational inference (VI) to tackle the problem of computational complexity, terming the model as a doubly sparse variational Gaussian process (S^2VGP).

Stochastic differential equations (SDEs) (Särkkä & Solin, 2019) provide

a framework to model complex dynamic systems with applications in diverse fields (Wang, 2005). An SDE consists of a deterministic term, drift, and a stochastic term, diffusion. Inference of an SDE involves learning of both of these functions. In many applications, the drift and diffusion are predefined from which inferring the SDE has been researched (Friedrich et al., 2011).

Ruttor et al. (2013); García (2017) research on learning the non-parametric drift and diffusion function using the Bayesian framework. These models result in intractable state distributions and thus use gradient matching algorithms. With the assumption of the drift function being linear, state distribution becomes Gaussian which opens the area for variational algorithms. Archambeau et al. (2007, 2008) research on a linear time-varying SDE method which on performing variational approximation provides an ELBO that is optimized using constrained optimization. As this work is closely related to the thesis, they are discussed in detail in Section 3.2.

Duncker et al. (2019) propose an extension of Archambeau et al. (2007). They propose to condition a GP on the drift function of the approximating SDE, and a mean-field variational approximation is performed between the drift and the state trajectory. In Ryder et al. (2018), the variational posterior is parameterized as an SDE whose drift is a neural network, and a discretized sampling scheme is used to evaluate the variational objective.

The advancement in automatic differentiation packages have encouraged researchers to explore the possibility of black-box learning of continuous-time dynamics. The most prominent work being neural ordinary differential equations (e.g. Chen et al., 2018; Rubanova et al., 2019). Li et al. (2020) introduced an efficient way to parameterize both the prior and posterior processes as SDE in a variational setting. Their method extends on Chen et al. (2018) and introduces a way to backpropagate through the SDE solution. It is general but requires discretization of the time horizon and cannot use adaptive SDE solvers.

Variational inference (VI) in models containing both the conjugate and non-conjugate terms are computationally very expensive. ELBO can be optimized using the stochastic-gradient methods; however, they might result in slow convergence as they do not use the conjugate properties. Khan (2014) proposed an efficient method to optimize the lower bound in variational approximation models termed as conjugate-computation variational inference (CVI).

3.2 Gaussian process approximations of stochastic differential equations

This section discusses the two research papers, Archambeau et al. (2007, 2008), in detail as they are closely related to the primary idea of the thesis. The paper presents an approach to perform Gaussian approximation to the posterior over paths for an SDE with observations. Following this, throughout the thesis, this method is referred to as GP-SDE.

Generative model

Suppose, there is a set of noisy observation $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$ corresponding to the hidden state $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$. An assumption about the prior (Itô) SDE is

$$\mathrm{d}\mathbf{x}(t) = \mathrm{f}\left(\mathbf{x}(t)\right)\ \mathrm{d}t + \sqrt{\mathbf{L}}\,\mathrm{d}\boldsymbol{\beta}(t). \tag{3.1}$$

In the presence of observations $\mathbf{Y}$, the posterior over paths is written as

$$\mathrm{d}p_{post}(\mathbf{x}(\cdot)) = \frac{1}{Z} \times \mathrm{d}p_{prior}(\mathbf{x}(\cdot)) \times \textstyle\prod_{i=1}^N p(\mathbf{y}_i \mid \mathbf{x}(t_i)), \tag{3.2}$$

where Z is the normalization constant and observations are over discrete time, paths being continuous. It is further assumed that the likelihood model is Gaussian, that is

$$p(\mathbf{y}_i \mid \mathbf{x}(t_i)) = \mathcal{N}\left(\mathbf{y}_i \mid \mathbf{H}\,\mathbf{x}(t_i), \mathbf{R}\right). \tag{3.3}$$

In many practical applications, discretization is favored, thus Eq. (3.1) using Euler–Maruyama with h time-step is written as

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k + \mathrm{f}(\mathbf{x}_k)\,h + \sqrt{h}\,\mathcal{N}(0,\,\mathbf{L}), \\ \delta\mathbf{x}_k = \mathbf{x}_{k+1} - \mathbf{x}_k &= \mathrm{f}(\mathbf{x}_k)\,h + \sqrt{h}\,\mathcal{N}(0,\,\mathbf{L}), \end{aligned} \tag{3.4}$$

where discretization is done over $[0, T]$, $\{k\}_{i=0}^{Th}$ is the grid index, and for brevity $\mathbf{x}(t_k)$ is written as $\mathbf{x}_k$. Therefore, in discrete space, t and $\mathbf{x}$ takes $[t_0, t_1, \ldots, t_{Th}]$ and $[\mathbf{x}_0,\, \mathbf{x}_1, \ldots,\, \mathbf{x}_{Th}]$ values respectively.

Approximate inference

In presence of observations, calculating the posterior is intractable thus Archambeau et al. (2007) employ approximate inference to obtain the same. In particular, VI, Section 2.2, is used to approximate the posterior over paths.

From the prior SDE, it is known that the process is Markovian. With an assumption of the drift function being linear, Archambeau et al. (2007) use a Markovian GP to approximate the posterior,

$$\mathrm{d}\mathbf{x}(t) = \mathrm{f}_{\mathrm{L}}(\mathbf{x}(t), t)\,\mathrm{d}t + \sqrt{\mathbf{L}}\,\mathrm{d}\boldsymbol{\beta}(t), \tag{3.5}$$

where $\mathrm{f}_{\mathrm{L}}(\mathbf{x}(t), t) = -\mathbf{A}(t)\,\mathbf{x}(t) + \mathbf{b}(t)$. Same diffusion term as the prior $\sqrt{\mathbf{L}}$ is used because of the KL divergence calculation which otherwise leads to infinity.

KL divergence calculation

Next, to compare the two measures, the approximating posterior q and the true posterior dp_{post} Archambeau et al. (2007) use KL divergence, Section 2.2.

For simplification, first the KL between prior SDE and the approximating posterior for discrete time is done. Employing Euler–Maruyama to the prior SDE, Eq. (3.1), gives

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{x}_k + \mathrm{f}(\mathbf{x}_k)\,h + \sqrt{h}\,\mathcal{N}(0,\,\mathbf{L}) \\
&= \mathcal{N}\left(\mathbf{x}_k + \mathrm{f}\left(\mathbf{x}_k\right)h, \mathbf{L}\,h\right), \\
p(\mathbf{x}_{0:K}) &= p\left(\mathbf{x}_0\right)\textstyle\prod_{i=1}^{K-1}\mathcal{N}\left(\mathbf{x}_k + \mathrm{f}\left(\mathbf{x}_k\right)h, \mathbf{L}\,h\right).
\end{aligned} \tag{3.6}$$

Similarly, for the approximating SDE, Eq. (3.5),

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{x}_k + \mathrm{f}_{\mathrm{L}}(\mathbf{x}_k,\,t_k)\,h + \sqrt{h}\,\mathcal{N}(0,\,\mathbf{L}) \\
&= \mathcal{N}(\mathbf{x}_k + \mathrm{f}_{\mathrm{L}}(\mathbf{x}_k,\,t_k)\,h,\;\mathbf{L}\,h), \\
q(\mathbf{x}_{0:K}) &= q(\mathbf{x}_0)\textstyle\prod_{k=1}^{K-1}\mathcal{N}(\mathbf{x}_k + \mathrm{f}_{\mathrm{L}}(\mathbf{x}_k,\,t_k)\,h, \mathbf{L}\,h).
\end{aligned} \tag{3.7}$$

Thus, the KL between them is evaluated as

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}_{0:K}) \,\|\, p(\mathbf{x}_{0:K})\right] &= \textstyle\int q(\mathbf{x}_{0:K}) \log \frac{q(\mathbf{x}_{0:K})}{p(\mathbf{x}_{0:K})}\,\mathrm{d}\mathbf{x}_{0:K} \\
&= \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}_0) \,\|\, p(\mathbf{x}_0)\right] + \textstyle\int q(\mathbf{x}_{1:K}) \log \frac{q(\mathbf{x}_{1:K})}{p(\mathbf{x}_{1:K})}\,\mathrm{d}\mathbf{x}_{1:K}.
\end{aligned} \tag{3.8}$$

Using the Markov property, the second term of the KL is further evaluated as

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}_{1:K}) \,\|\, p(\mathbf{x}_{1:K})\right] = \textstyle\sum_{i=1}^{K-1}\int q(\mathbf{x}_i)\,\mathrm{d}\mathbf{x}_i \int q(\mathbf{x}_{i+1} \mid \mathbf{x}_i) \log \frac{q(\mathbf{x}_{i+1}|\mathbf{x}_i)}{p(\mathbf{x}_{i+1}|\mathbf{x}_i)}\,\mathrm{d}\mathbf{x}_{i+1} \\
&= \textstyle\sum_{i=1}^{K-1}\int q(\mathbf{x}_i)\,\mathrm{d}\mathbf{x}_i \int \mathrm{d}\mathbf{x}_{i+1}\, q(\mathbf{x}_{i+1} \mid \mathbf{x}_i)\,\left[\log q(\mathbf{x}_{i+1} \mid \mathbf{x}_i) - \log p(\mathbf{x}_{i+1} \mid \mathbf{x}_i)\right].
\end{aligned} \tag{3.9}$$

Next, for simplicity the log terms are evaluated separately,

$$\begin{aligned}
&\log q(\mathbf{x}_{i+1} \mid \mathbf{x}_i) - \log p(\mathbf{x}_{i+1} \mid \mathbf{x}_i) \\
&\quad = -\frac{1}{2} \left(\mathbf{x}_{i+1} - \mathbf{x}_i - h\, \mathrm{f_L}(\mathbf{x}_i, t_i)\right)^\top (\mathbf{L}\, h)^{-1} \left(\mathbf{x}_{i+1} - \mathbf{x}_i - h\, \mathrm{f_L}(\mathbf{x}_i, t_i)\right) \\
&\qquad + \frac{1}{2} \left(\mathbf{x}_{i+1} - \mathbf{x}_i - h\, \mathrm{f}(\mathbf{x}_i)\right)^\top (\mathbf{L}\, h)^{-1} \left(\mathbf{x}_{i+1} - \mathbf{x}_i - h\, \mathrm{f}(\mathbf{x}_i)\right) \\
&\quad = -\frac{\mathbf{L}^{-1}}{2} \big(-\Delta\mathbf{x}_i^\top\, \mathrm{f_L}(\mathbf{x}_i, t_i) - \mathrm{f_L}(\mathbf{x}_i, t_i)^\top \Delta\mathbf{x}_i + \mathrm{f_L}(\mathbf{x}_i, t_i)^\top h\, \mathrm{f_L}(\mathbf{x}_i, t_i) \\
&\qquad + \Delta\mathbf{x}_i^\top \mathrm{f}(\mathbf{x}_i) + \mathrm{f}(\mathbf{x}_i)^\top \Delta\mathbf{x}_i - \mathrm{f}(\mathbf{x}_i)^\top h\, \mathrm{f}(\mathbf{x}_i)\big)\,. \qquad (3.10)
\end{aligned}$$

Further, it is known that $\mathbb{E}_{q(\mathbf{x}_i)}[\Delta\mathbf{x}_i \mid \mathbf{x}_i] = \mathrm{f_L}(\mathbf{x}_i, t_i)\, h$. Therefore,

$$\begin{aligned}
\mathrm{D_{KL}}\left[q(\mathbf{x}_{1:K}) \,\|\, p(\mathbf{x}_{1:K})\right] =& \frac{1}{2} \textstyle\sum_{i=1}^{K-1} \int \mathrm{d}\mathbf{x}_i\, q(\mathbf{x}_i)\, h \big[(\mathrm{f}(\mathbf{x}_i) - \mathrm{f_L}(\mathbf{x}_i, t_i))^\top \mathbf{L}^{-1} (\mathrm{f}(\mathbf{x}_i) \\
& - \mathrm{f_L}(\mathbf{x}_i, t_i))\big], \qquad (3.11)
\end{aligned}$$

which leads to

$$\begin{aligned}
\mathrm{D_{KL}}\left[q(\mathbf{x}_{0:K}) \,\|\, p(\mathbf{x}_{0:K})\right] &= \mathrm{D_{KL}}\left[q(\mathbf{x}_0) \,\|\, p(\mathbf{x}_0)\right] \\
&\quad + \frac{1}{2} \textstyle\sum_{i=1}^{K-1} h \left\langle (\mathrm{f}(\mathbf{x}_i) - \mathrm{f_L}(\mathbf{x}_i, t_i))^\top \mathbf{L}^{-1} (\mathrm{f}(\mathbf{x}_i) - \mathrm{f_L}(\mathbf{x}_i, t_i)) \right\rangle_{q(\mathbf{x}_i)}, \qquad (3.12)
\end{aligned}$$

where $\langle \cdot \rangle_{q(\mathbf{x}_i)}$ means expectation under the posterior distribution q at $\mathbf{x}_i$.

As the terms have linear scaling with h of Riemann sums, the final KL is written as

$$\begin{aligned}
\mathrm{D_{KL}}\left[q(\mathbf{x}) \,\|\, p(\mathbf{x})\right] &= \mathrm{D_{KL}}\left[q(\mathbf{x}_0) \,\|\, p(\mathbf{x}_0)\right] \\
&\quad + \frac{1}{2} \textstyle\int_i \left\langle (\mathrm{f}(\mathbf{x}_i) - \mathrm{f_L}(\mathbf{x}_i,\, t_i))^\top \mathbf{L}^{-1} (\mathrm{f}(\mathbf{x}_i) - \mathrm{f_L}(\mathbf{x}_i,\, t_i)) \right\rangle_{q(\mathbf{x}_i)} \mathrm{d}t_i. \qquad (3.13)
\end{aligned}$$

Similarly, KL between the approximating and true posterior is calculated as

$$\begin{aligned}
\mathrm{D_{KL}}\left[q(\mathbf{x}) \,\|\, p_{post}(\mathbf{x})\right] &= \textstyle\int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p_{post}(\mathbf{x})}\, \mathrm{d}\mathbf{x} \\
&= \textstyle\int \mathrm{d}\mathbf{x}\, q(\mathbf{x}) \log q(\mathbf{x}) - q(\mathbf{x}) \log p_{post}(\mathbf{x}) \\
&= \textstyle\int \mathrm{d}\mathbf{x}\, q(\mathbf{x}) \log q(\mathbf{x}) - q(\mathbf{x}) \big[- \log Z + \log p_{sde}(\mathbf{x}) + \\
&\qquad \textstyle\sum_{i=1}^{N} \log \mathcal{N}(\mathbf{y}_i \mid \mathbf{H}\, \mathbf{x}(t_i),\, \mathbf{R})\big] \\
&= \textstyle\int \mathrm{d}\mathbf{x}\, q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p_{sde}(\mathbf{x})} + q(\mathbf{x}) \log Z - q(\mathbf{x}) \big[\sum_{i=1}^{N} -\frac{D}{2} \log 2\pi \\
&\quad - \frac{1}{2} \log |\mathbf{R}| - \frac{1}{2} (\mathbf{y}_i - \mathbf{H}\, \mathbf{x}(t_i))^\top \mathbf{R}^{-1} (\mathbf{y}_i - \mathbf{H}\, \mathbf{x}(t_i))\big]
\end{aligned}$$

$$= \log Z + \frac{N\,D}{2}\log 2\pi + \frac{N}{2}\log|\mathbf{R}| + \int \mathrm{d}\mathbf{x}\; q(\mathbf{x}) \log \tfrac{q(\mathbf{x})}{p_{sde}(\mathbf{x})} + \frac{1}{2}\sum_{i=1}^{N} \int \mathrm{d}\mathbf{x}\; q(\mathbf{x})(\mathbf{y}_i - \mathbf{H}\,\mathbf{x}(t_i))^\top \mathbf{R}^{-1}(\mathbf{y}_i - \mathbf{H}\,\mathbf{x}(t_i)). \tag{3.14}$$

Thus, the final KL is

$$\mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}) \,\|\, p_{post}(\mathbf{x})\right] = \log Z + \frac{ND}{2}\log 2\pi + \frac{N}{2}\log|\mathbf{R}| + \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}_0) \,\|\, p(\mathbf{x}_0)\right] + \int_t E_{sde}(t) + E_{obs}(t)\,\mathrm{d}t, \tag{3.15}$$

where
$E_{sde}(t) = \frac{1}{2}\left\langle (\mathrm{f}(\mathbf{x}_t) - \mathrm{f_L}(\mathbf{x}_t,\, t))^\top \mathbf{L}^{-1} (\mathrm{f}(\mathbf{x}_t) - \mathrm{f_L}(\mathbf{x}_t,\, t)) \right\rangle_{q(\mathbf{x}_t)}$,
$E_{obs}(t) = \frac{1}{2}\sum_{i=1}^{N} \left\langle (\mathbf{y}_i - \mathbf{H}\,\mathbf{x}(t_i))^\top \mathbf{R}^{-1} (\mathbf{y}_i - \mathbf{H}\,\mathbf{x}(t_i)) \right\rangle_{q(\mathbf{x}_t)} \delta(t - t_i)$,

and $\delta(\cdot)$ is the dirac function.

Moments of the approximating posterior

The approximation posterior is given by the SDE,

$$\mathrm{d}\mathbf{x}(t) = \mathrm{f_L}(\mathbf{x}(t),\, t)\,\mathrm{d}t + \sqrt{\mathbf{L}}\,\mathrm{d}\boldsymbol{\beta}(t)\,,$$

and, as the SDE is linear, which implies Gaussianity, the posterior is

$$q(\mathbf{x}(t)) \sim \mathcal{N}(\mathbf{m}(t),\, \mathbf{S}(t)).$$

As discussed in Särkkä & Solin (2019, Chapter 6), a Gaussian distribution is completely defined by its two moments which are given by

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} &= \mathbb{E}\left[\mathrm{f_L}\left(\mathbf{x}(t),\, t\right)\right] = -\mathbf{A}(t)\,\mathbf{m}(t) + \mathbf{b}(t), \\
\frac{\mathrm{d}\mathbf{S}}{\mathrm{d}t} &= \mathbb{E}[\mathrm{f_L}(\mathbf{x}(t),\, t)\,(\mathbf{x}(t) - \mathbf{m}(t))^\top] + \mathbb{E}[(\mathbf{x}(t) - \mathbf{m}(t))\,\mathrm{f_L}^\top(\mathbf{x}(t),\, t)] \\
&\quad + \mathbb{E}[\sqrt{\mathbf{L}}\,\mathbf{Q}\,\sqrt{\mathbf{L}}^\top] \\
&= -\mathbf{A}(t)\,\mathbf{S}(t) - \mathbf{S}(t)\,\mathbf{A}^\top(t) + \mathbf{L}.
\end{aligned} \tag{3.16}$$

Therefore, the two moments are available by solving the system

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} &= -\mathbf{A}(t)\,\mathbf{m}(t) + \mathbf{b}(t), \\
\frac{\mathrm{d}\mathbf{S}}{\mathrm{d}t} &= -\mathbf{A}(t)\,\mathbf{S}(t) - \mathbf{S}(t)\,\mathbf{A}^\top(t) + \mathbf{L}.
\end{aligned} \tag{3.17}$$

Learning the approximate posterior

The aim is to learn the approximate posterior such that the KL divergence between the approximating posterior and the true posterior Eq. (3.15) is minimum. Also, the two moments of the posterior are given by Eq. (3.17).

Archambeau et al. (2007) perform constrained optimization (Bertsekas, 1996) where the main objective is to minimize the KL divergence constrained on the two moments using Lagrange multipliers,

$$\ell = \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}_0) \,\|\, p(\mathbf{x}_0)\right] + \int_{t_0}^{T} E(t) - \mathrm{tr}\left\{\boldsymbol{\Psi}(t)\left(\tfrac{\mathrm{d}\mathbf{S}}{\mathrm{d}t} + \mathbf{A}(t)\,\mathbf{S}(t) + \mathbf{S}(t)\,\mathbf{A}^\top(t) - \mathbf{L}\right)\right\} - \boldsymbol{\lambda}(t)^\top\left(\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} + \mathbf{A}(t)\,\mathbf{m}(t) - \mathbf{b}(t)\right)\mathrm{d}t, \tag{3.18}$$

where $E(t) = E_{obs}(t) + E_{sde}(t)$.

Applying integration by parts leads to

$$\begin{aligned}\ell &= \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}_0) \,\|\, p(\mathbf{x}_0)\right] + \int_{t_0}^{T} E(t) - \mathrm{tr}\left\{\boldsymbol{\Psi}(t)(\mathbf{A}(t)\,\mathbf{S}(t) + \mathbf{S}(t)\,\mathbf{A}^\top(t) - \mathbf{L})\right\} \\ &\quad - \boldsymbol{\lambda}(t)^\top(\mathbf{A}(t)\,\mathbf{m}(t) - \mathbf{b}(t)) - \mathrm{tr}\left\{\boldsymbol{\Psi}(t)\frac{\mathrm{d}\mathbf{S}}{\mathrm{d}t}\right\} - \boldsymbol{\lambda}(t)^\top\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t}\,\mathrm{d}t \\ &= \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}_0) \,\|\, p(\mathbf{x}_0)\right] + \int_{t_0}^{T} E(t) - \mathrm{tr}\left\{\boldsymbol{\Psi}(t)(\mathbf{A}(t)\,\mathbf{S}(t) + \mathbf{S}(t)\,\mathbf{A}^\top(t) - \mathbf{L})\right\} \\ &\quad - \boldsymbol{\lambda}(t)^\top(\mathbf{A}(t)\,\mathbf{m}(t) - \mathbf{b}(t)) + \mathrm{tr}\left\{\mathbf{S}\,\frac{\mathrm{d}\boldsymbol{\Psi}(t)}{\mathrm{d}t}\right\} + \mathbf{m}(t)\,\frac{\mathrm{d}\boldsymbol{\lambda}^\top(t)}{\mathrm{d}t}\,\mathrm{d}t \\ &\quad - \mathrm{tr}\{\boldsymbol{\Psi}(T)\,\mathbf{S}(T)\} + \mathrm{tr}\{\boldsymbol{\Psi}(0)\,\mathbf{S}(0)\} - \boldsymbol{\lambda}^\top(T)\,\mathbf{m}(T) + \boldsymbol{\lambda}^\top(0)\,\mathbf{m}(0).\end{aligned} \tag{3.19}$$

To optimize and update the values of the parameters $\mathbf{A}$, $\mathbf{b}$, $\mathbf{m}$, $\mathbf{S}$, Archambeau et al. (2007) suggest calculating the gradient of the objective, Eq. (3.19), with respect to them and setting them to zero,

$$\begin{aligned}\frac{\partial\ell(t)}{\partial\mathbf{A}(t)} &= \frac{\partial E(t)}{\partial\mathbf{A}(t)} - 2\boldsymbol{\Psi}(t)\,\mathbf{S}(t) - \boldsymbol{\lambda}(t)\,\mathbf{m}(t)^\top = 0, \\ \frac{\partial\ell(t)}{\partial\mathbf{b}(t)} &= \frac{\partial E(t)}{\partial\mathbf{b}(t)} + \boldsymbol{\lambda}(t) = 0, \\ \frac{\partial\ell(t)}{\partial\mathbf{S}(t)} &= \frac{\partial E(t)}{\partial\mathbf{S}(t)} - 2\boldsymbol{\Psi}(t)\,\mathbf{A}(t) + \frac{\mathrm{d}\boldsymbol{\Psi}(t)}{\mathrm{d}t} = 0, \\ \frac{\partial\ell(t)}{\partial\mathbf{m}(t)} &= \frac{\partial E(t)}{\partial\mathbf{m}(t)} - \mathbf{A}(t)^\top\,\boldsymbol{\lambda}(t) + \frac{\mathrm{d}\boldsymbol{\lambda}(t)}{\mathrm{d}t} = 0.\end{aligned} \tag{3.20}$$

Further, using the property $E(t) = E_{sde}(t) + E_{obs}(t)$,

$$\frac{\partial E(t)}{\partial\mathbf{A}(t)} = \frac{\partial E_{sde}(t)}{\partial\mathbf{A}(t)}$$

$$= \mathbf{L}^{-1}\left[\left\langle \mathrm{f}(\mathbf{x}(t))\,\mathbf{x}(t)^\top\right\rangle_{q(\mathbf{x}_t)} + \mathbf{A}(t)\left\langle \mathbf{x}(t)^\top\mathbf{x}(t)\right\rangle_{q(\mathbf{x}_t)} - \mathbf{b}(t)\left\langle \mathbf{x}(t)^\top\right\rangle_{q(\mathbf{x}_t)}\right], \tag{3.21}$$

$$\frac{\partial E(t)}{\partial \mathbf{b}(t)} = \frac{\partial E_{sde}(t)}{\partial \mathbf{b}(t)}$$
$$= -\mathbf{L}^{-1}\left[\left\langle \mathrm{f}(\mathbf{x}(t))\right\rangle_{q(\mathbf{x}_t)} + \mathbf{A}(t)\left\langle \mathbf{x}(t)\right\rangle_{q(\mathbf{x}_t)} - \mathbf{b}(t)\right]. \tag{3.22}$$

Further using the properties, $\left\langle \mathbf{x}\,\mathbf{x}^\top\right\rangle_{q(\mathbf{x}_t)} = \mathbf{m}\,\mathbf{m}^\top + \mathbf{S}$ and $\left\langle \mathrm{f}(\mathbf{x})\,(\mathbf{x}-\mathbf{m})^\top\right\rangle_{q(\mathbf{x}_t)} = \left\langle \frac{\partial \mathrm{f}}{\partial \mathbf{x}}\right\rangle_{q(\mathbf{x}_t)}\mathbf{S}$, Archambeau et al. (2007) write the update rules as

$$\begin{aligned}
\tilde{\mathbf{A}}(t) &= -\left\langle \frac{\partial \mathrm{f}}{\partial \mathbf{x}}\right\rangle_{q(\mathbf{x}_t)} + 2\,\mathbf{L}\,\mathbf{\Psi}(t),\\
\tilde{\mathbf{b}}(t) &= \left\langle \mathrm{f}(\mathbf{x})\right\rangle_{q(\mathbf{x}_t)} + \tilde{\mathbf{A}}(t)\,\mathbf{m}(t) - \mathbf{L}\,\lambda(t),\\
\frac{\mathrm{d}\mathbf{\Psi}(t)}{\mathrm{d}t} &= 2\,\mathbf{\Psi}(t)\,\mathbf{A}(t) - \frac{\partial E(t)}{\partial \mathbf{S}(t)},\\
\frac{\mathrm{d}\boldsymbol{\lambda}(t)}{\mathrm{d}t} &= \mathbf{A}(t)^\top\boldsymbol{\lambda}(t) - \frac{\partial E(t)}{\partial \mathbf{m}(t)}.
\end{aligned} \tag{3.23}$$

To update the Lagrange multipliers, when an observation is present, a jump condition is performed (derivative of E_{obs} tells amplitude)

$$\mathbf{\Psi}(t_n^+) = \mathbf{\Psi}(t_n^-) - \frac{1}{2}\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}, \tag{3.24}$$

$$\boldsymbol{\lambda}(t_n^+) = \boldsymbol{\lambda}(t_n^-) + \mathbf{H}^\top\mathbf{R}^{-1}(\mathbf{y}_n - \mathbf{H}\,\mathbf{m}(t_n)). \tag{3.25}$$

However, due to numerical stability, Archambeau et al. (2007) suggest updating the parameter values of $\mathbf{A}$, $\mathbf{b}$ by taking small steps

$$\mathbf{A}(t) = \mathbf{A}(t) - \omega(\mathbf{A}(t) - \tilde{\mathbf{A}}(t)), \tag{3.26}$$

$$\mathbf{b}(t) = \mathbf{b}(t) - \omega(\mathbf{b}(t) - \tilde{\mathbf{b}}(t)), \tag{3.27}$$

where $\omega \in (0, 1)$.

Similarly, to learn and optimize the initial states, gradient of the objective, Eq. (3.19), with respect to $\mathbf{m}_0$, $\mathbf{S}_0$ is calculated and set to zero. The prior distribution on the initial state $p(\mathbf{x}_0)$ is chosen to be a Gaussian that is $p(\mathbf{x}_0) \sim \mathcal{N}(\mathbf{m}_p, \mathbf{S}_p)$. Thus, the KL divergence between the posterior and prior over initial states is written as

$$\mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}_0)\,\|\,p(\mathbf{x}_0)\right] = \frac{1}{2}\left[\log\frac{|\mathbf{S}_p|}{|\mathbf{S}_0|} - k + (\mathbf{m}_0 - \mathbf{m}_p)^\top\mathbf{S}_p^{-1}(\mathbf{m}_0 - \mathbf{m}_p)\right.$$

$$+ \operatorname{tr}\left\{\mathbf{S}_p^{-1}\mathbf{S}_0\right\}\Big]\,. \tag{3.28}$$

The derivative of Eq. (3.28) with respect to $\mathbf{m}_0$ and $\mathbf{S}_0$ is

$$\frac{\partial\, \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}_0)\,\|\,p(\mathbf{x}_0)\right]}{\partial \mathbf{m}_0} = \mathbf{S}_p^{-1}\left(\mathbf{m}_0 - \mathbf{m}_p\right), \tag{3.29}$$

$$\frac{\partial\, \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}_0)\,\|\,p(\mathbf{x}_0)\right]}{\partial \mathbf{S}_0} = \frac{1}{2}\left[\mathbf{S}_p^{-1} - \mathbf{S}_0^{-1}\right]. \tag{3.30}$$

Thus, using the above result, the gradients of the objective Eq. (3.19) leads to

$$\frac{\partial \ell}{\partial\, \mathbf{m}_0} = \mathbf{S}_p^{-1}\left(\mathbf{m}_0 - \mathbf{m}_p\right) + \boldsymbol{\lambda}(0)\,, \tag{3.31}$$

$$\frac{\partial \ell}{\partial\, \mathbf{S}_0} = \frac{1}{2}\left[\mathbf{S}_p^{-1} - \mathbf{S}_0^{-1}\right] + \boldsymbol{\Psi}(0)\,, \tag{3.32}$$

and by setting them to zero the update rule is

$$\mathbf{m}_0 = \mathbf{m}_p - \mathbf{S}_p\,\boldsymbol{\lambda}(0), \tag{3.33}$$

$$\mathbf{S}_0 = \left[\mathbf{S}_p^{-1} + 2\,\boldsymbol{\Psi}(0)\right]^{-1}. \tag{3.34}$$

This method is also implemented and experimented with the two processes: Ornstein–Uhlenbeck (OU) and double-well process in Chapter 5.

Chapter 4

Methods

This chapter introduces the proposed method, derives the variational inference objective that is evidence lower bound (ELBO) using Girsanov's theorem. An optimization algorithm based on natural gradients is also presented to learn the model parameters. Following this, throughout the thesis, the proposed method is referred to as SGP-SDE.

4.1 Sparse Markovian process

Suppose, a latent process is governed by a one-dimensional state SDE p that is represented by its drift function ($f = f_\theta$) and diffusion term ($L = \sqrt{\Sigma}$) as

$$p_\theta(\mathbf{x}) : \ \mathrm{d}\mathbf{x}(t) = f_\theta(\mathbf{x}(t))\mathrm{d}t + \sqrt{\Sigma}\,\mathrm{d}\boldsymbol{\beta}(t), \tag{4.1}$$

and $\mathbf{y}$ is observed through an observation model. The proposed idea is to approximate the posterior of this process $p_\theta(\mathbf{X} \mid \mathbf{Y})$ by a sparse Markovian Gaussian process q. As discussed in Section 2.6, the posterior of q is written as

$$q_{\{\psi,\xi\}}(\mathbf{x}(\cdot),\, \mathbf{x}(\mathbf{z})) = r_\psi(\mathbf{x}(\cdot) \mid \mathbf{x}(\mathbf{z}))\; w_\xi(\mathbf{x}(\mathbf{z})), \tag{4.2}$$

where $r_\psi(\mathbf{x}(\cdot) \mid \mathbf{x}(\mathbf{z}))$ is a probability density of a Gaussian process (GP) and $w_\xi(\mathbf{x}(\mathbf{z}))$ is a probability density over the inducing variables $\mathbf{x}(\mathbf{z})$. It should be noted that $\mathbf{x}(\cdot)$ are the values of $\mathbf{x}$ everywhere including the inducing locations $\mathbf{z}$ thus the joint distribution is written as
$q_{\{\psi,\xi\}}(\mathbf{x}(\cdot)) = q_{\{\psi,\xi\}}(\mathbf{x}(\cdot),\, \mathbf{x}(\mathbf{z}))$.

As discussed in Särkkä & Solin (2019, Chapter 12), the Markovian GP r_ψ is expressed as a linear time invariant SDE

$$r_\psi(\mathbf{x}) : \ \mathrm{d}\mathbf{x}(t) = \mathbf{F}_\psi\, \mathbf{x}\, \mathrm{d}t + \sqrt{\Sigma}\,\mathrm{d}\boldsymbol{\beta}(t), \tag{4.3}$$

where the diffusion term is same as the specifying process p which otherwise would lead to infinity in the following calculations.

For approximate inference, variational inference is used where objective is to maximize the ELBO. As part of the ELBO computation, KL needs to be calculated between the true and the approximated process

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}(\cdot)) \,\|\, p(\mathbf{x}(\cdot))\right] &= \mathbb{E}_{q(\mathbf{x}(\cdot))}\left[\log \frac{q(\mathbf{x}(\cdot))}{p(\mathbf{x}(\cdot))}\right] \\
&= \mathbb{E}_{q(\mathbf{x}(\cdot))}\left[\log \frac{r(\mathbf{x}(\cdot) \mid \mathbf{x}(\mathbf{z}))\, w(\mathbf{x}(\mathbf{z}))}{p(\mathbf{x}(\cdot))}\right] \\
&= \mathbb{E}_{q(\mathbf{x}(\cdot))}\left[\log \frac{r(\mathbf{x}(\cdot) \mid \mathbf{x}(\mathbf{z}))\, r(\mathbf{x}(\mathbf{z}))}{p(\mathbf{x}(\cdot))} + \log \frac{w(\mathbf{x}(\mathbf{z}))}{r(\mathbf{x}(\mathbf{z}))}\right] \\
&= -\,\mathbb{E}_{q(\mathbf{x}(\cdot))}\left[\log \frac{p(\mathbf{x}(\cdot))}{r(\mathbf{x}(\cdot))}\right] + \mathbb{E}_{q(\mathbf{x}(\mathbf{z}))}\left[\log \frac{w(\mathbf{x}(\mathbf{z}))}{r(\mathbf{x}(\mathbf{z}))}\right],
\end{aligned} \tag{4.4}$$

where the first term is continuous over paths (infinite dimensional term) and the second term is discrete over the inducing variables z (finite dimensional term). Both the terms are evaluated separately.

Infinite dimensional term

The infinite dimensional term, that is the continuous over paths term in Eq. (4.4), is evaluated using Girsanov's theorem (Girsanov, 1960).

Girsanov's theorem: Särkkä & Sottinen (2008) further derived the Girsanov theorem to provide a way to calculate the likelihood ratio of two Itô processes. Consider, the two Itô processes are

$$\begin{aligned}
p(\mathbf{x}) : \mathrm{d}\mathbf{x} &= f(\mathbf{x},\, t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}, \\
q(\mathbf{y}) : \mathrm{d}\mathbf{y} &= \mathrm{g}(\mathbf{y},\, t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\beta},
\end{aligned}$$

with both the Brownian motions having $\mathbf{Q}$ spectral density.
The likelihood ratio between the two processes can be written as

$$\begin{aligned}
\frac{p(\mathbf{X})}{p(\mathbf{Y})} = \exp\Bigg(&-\frac{1}{2}\int_{t=0}^{\tau} [f(\mathbf{y},\, t) - \mathrm{g}(\mathbf{y}, t)]^{\top} \mathbf{Q}^{-1}\, [f(\mathbf{y},\, t) - \mathrm{g}(\mathbf{y},\, t)]\,\mathrm{d}t \\
&+ \int_{t=0}^{\tau} [f(\mathbf{y},\, t) - \mathrm{g}(\mathbf{y},\, t)]^{\top} \mathbf{Q}^{-1}\,\mathrm{d}\boldsymbol{\beta}(t)\Bigg).
\end{aligned} \tag{4.5}$$

Following Eq. (4.5), the infinite dimensional term is evaluated as

$$\mathbb{E}_{q(\mathbf{x}(\cdot))}\left[\log \frac{p(\mathbf{x}(\cdot))}{r(\mathbf{x}(\cdot))}\right] = \mathbb{E}_{q(\mathbf{x}(\cdot))}\Bigg[-\frac{1}{2}\int_{t=0}^{\tau}[f(\mathbf{x}_t)-\mathbf{F}\,\mathbf{x}_t]^\top \Sigma^{-1}\,[f(\mathbf{x}_t)-\mathbf{F}\,\mathbf{x}_t]\,\mathrm{d}t + \int_{t=0}^{\tau}[f(\mathbf{x}_t)-\mathbf{F}\,\mathbf{x}_t]^\top \Sigma\,\mathrm{d}\boldsymbol{\beta}(t)\Bigg], \tag{4.6}$$

where for brevity $\mathbf{x}(t)$ is written as $\mathbf{x}_t$. Further, the second term in the above equation is zero as

$$\begin{aligned}\mathbb{E}_{q(\mathbf{x}(\cdot))}\left[\int_{t=0}^{\tau}[f(\mathbf{x}_t)-\mathbf{F}\,\mathbf{x}_t]^\top \Sigma\,\mathrm{d}\boldsymbol{\beta}(t)\right] &= \mathbb{E}_{q(\mathbf{x}(\cdot))}\left[\phi(\mathbf{X})\,\mathrm{d}\boldsymbol{\beta}(t)\right]\\ &= 0\,,\end{aligned} \tag{4.7}$$

where $\phi(\mathbf{X})$ is a transformation of $\mathbf{X}$. Thus, the infinite dimensional term is evaluated as

$$\mathbb{E}_{q(\mathbf{x}(\cdot))}\left[\log \frac{r(\mathbf{x}(\cdot))}{p(\mathbf{x}(\cdot))}\right] = -\frac{1}{2}\int_{t=0}^{\tau}\mathbb{E}_{q(\mathbf{x}_t)}\left[(f(\mathbf{x}_t)-\mathbf{F}\,\mathbf{x}_t)^\top \Sigma^{-1}(f(\mathbf{x}_t)-\mathbf{F}\,\mathbf{x}_t)\right]\,\mathrm{d}t, \tag{4.8}$$

where $\mathbf{x}_t = \mathbf{x}(t)$.

Finite dimensional term

From the model definition Eq. (4.2), it is known that $q(\mathbf{x}(\mathbf{z}))$ and $w(\mathbf{x}(\mathbf{z}))$ are equal. Therefore, the finite term, that is the discrete term over inducing variables in Eq. (4.4), is evaluated as

$$\begin{aligned}\mathbb{E}_{q(\mathbf{x}(\mathbf{z}))}\left[\log \frac{w(\mathbf{x}(\mathbf{z}))}{r(\mathbf{x}(\mathbf{z}))}\right] &= \mathbb{E}_{w(\mathbf{x}(\mathbf{z}))}\left[\log \frac{w(\mathbf{x}(\mathbf{z}))}{r(\mathbf{x}(\mathbf{z}))}\right]\\ &= \mathrm{D}_{\mathrm{KL}}\left[w(\mathbf{x}(\mathbf{z}))\,\|\,r(\mathbf{x}(\mathbf{z}))\right].\end{aligned} \tag{4.9}$$

Therefore, the KL between p and the approximating process q using Eq. (4.8) and Eq. (4.9) is

$$\mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x})\,\|\,p(\mathbf{x})\right] = \frac{1}{2}\int_{t=0}^{\tau}\mathbb{E}_{q(\mathbf{x}_t)}\left[(f(\mathbf{x}_t)-\mathbf{F}\,\mathbf{x}_t)^\top \Sigma^{-1}\,(f(\mathbf{x}_t)-\mathbf{F}\,\mathbf{x}_t)\right]\,\mathrm{d}t + \mathrm{D}_{\mathrm{KL}}\left[w(\mathbf{x}(\mathbf{z}))\,\|\,r(\mathbf{x}(\mathbf{z}))\right]. \tag{4.10}$$

4.2 Evidence lower bound (ELBO)

As discussed in Section 2.2, the ELBO is a combination of the KL and the variational log-likelihood. Therefore, for the proposed model, the ELBO is

$$\ell = -\frac{1}{2}\int_{t=0}^{\tau}\mathbb{E}_{q(\mathbf{x}_t)}\left[(f(\mathbf{x}_t)-\mathbf{F}\,\mathbf{x}_t)^\top \Sigma^{-1}\,(f(\mathbf{x}_t)-\mathbf{F}\,\mathbf{x}_t)\right]\,\mathrm{d}t$$

$$- \mathrm{D}_{\mathrm{KL}}\left[w(\mathbf{x}(\mathbf{z})) \,\|\, r(\mathbf{x}(\mathbf{z}))\right] + \mathbb{E}_{q(\mathbf{x}(\cdot))}[\log p(\mathbf{Y} \mid \mathbf{X})]. \tag{4.11}$$

With an assumption that the observations are independent and identically distributed (IID), it is further written as

$$\begin{aligned}\ell = & -\frac{1}{2}\int_{t=0}^{\tau} \mathbb{E}_{q(\mathbf{x}_t)}\left[(f(\mathbf{x}_t) - \mathbf{F}\,\mathbf{x}_t)^\top \Sigma^{-1}\,(f(\mathbf{x}_t) - \mathbf{F}\,\mathbf{x}_t)\right]\,\mathrm{d}t \\ & - \mathrm{D}_{\mathrm{KL}}\left[w(\mathbf{x}(\mathbf{z})) \,\|\, r(\mathbf{x}(\mathbf{z}))\right] + \textstyle\sum_n \mathbb{E}_{q(\mathbf{x}_n)}[\log p(\mathbf{y}_n \mid \mathbf{x}_n)].\end{aligned} \tag{4.12}$$

For brevity, in the following sections, the ELBO is written as

$$\begin{aligned}\ell(\theta,\, \psi,\, \xi) = & \int_{t=0}^{\tau} \mathbb{E}_{q_\phi(\mathbf{x}_t)}\left[h_1(\mathbf{x}_t)\right]\,\mathrm{d}t + \textstyle\sum_n \mathbb{E}_{q_\phi(\mathbf{x}_n)}\left[h_2(\mathbf{x}_n)\right] \\ & - \mathrm{D}_{\mathrm{KL}}\left[w_\xi(\mathbf{x}_z) \,\|\, r_\psi(\mathbf{x}_z)\right],\end{aligned} \tag{4.13}$$

where $h_1(\mathbf{x}_t) = -\frac{1}{2}(f(\mathbf{x}_t) - \mathbf{F}\,\mathbf{x}_t)^\top \Sigma^{-1}\,(f(\mathbf{x}_t) - \mathbf{F}\,\mathbf{x}_t)$ and $h_2(\mathbf{x}_n) = \log p(\mathbf{y}_n \mid \mathbf{x}_n)$.

The terms of the ELBO are interpreted as the drift matching term being the first one, complexity being the second term, and third term managing the model fit.

KL term

From model definition, it is known that $q(\mathbf{x}_\mathbf{z}) = w(\mathbf{x}_\mathbf{z})$. Using this property along with Gaussian properties, KL term in Eq. (4.13) is evaluated as

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}\left[w(\mathbf{x}_\mathbf{z}) \,\|\, r(\mathbf{x}_\mathbf{z})\right] &= \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{x}_\mathbf{z}) \,\|\, r(\mathbf{x}_\mathbf{z})\right] \\ &= \mathbb{E}_{q(\mathbf{x}_\mathbf{z})}\left[\log q(\mathbf{x}_\mathbf{z})\right] - \mathbb{E}_{q(\mathbf{x}_\mathbf{z})}\left[\log r(\mathbf{x}_\mathbf{z})\right] \\ &= -\frac{1}{2}\log|\Sigma_{w_\mathbf{z}\,w_\mathbf{z}}| + \frac{1}{2}\mathrm{tr}(\Sigma_{w_\mathbf{z}\,w_\mathbf{z}}\,\mathbf{K}^{-1}) + const,\end{aligned} \tag{4.14}$$

where $w(\mathbf{x}_\mathbf{z}) \sim \mathcal{N}(\mu_{w_\mathbf{z}}, \Sigma_{w_\mathbf{z}\,w_\mathbf{z}})$ and $\mathbf{K}$ is the kernel of the GP r.

4.3 Natural gradient descent

Stochastic gradient descent (SGD) can be used to optimize the ELBO in Eq. (4.13) treating it as a black-box. However, as discussed by Manfred Opper (2009), it can be slow as the conjugate-computation benefits are not being utilized as well as the number of free parameters are more in SGD. Thus, natural gradients are used.

The objective is to maximize the ELBO in Eq. (4.13) with respect to the distribution $w(\mathbf{x}_\mathbf{z})$. As discussed earlier, the objective of maximizing the ELBO can also be termed as the minimization of the negative ELBO.

Let $w(\mathbf{x_z})$ be parameterized by $\boldsymbol{\lambda}$. By using SGD, the parameter of the distribution is optimized as

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \rho_t \nabla_{\boldsymbol{\lambda}} \mathfrak{L}(\boldsymbol{\lambda}_t). \tag{4.15}$$

However, natural gradients can also be used which provides better updates and is independent of the parameterization

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \rho_t \, \mathrm{F}(\boldsymbol{\lambda}_t)^{-1} \nabla_{\boldsymbol{\lambda}} \mathfrak{L}(\boldsymbol{\lambda}_t), \tag{4.16}$$

where $\mathrm{F}(\cdot)$ is the Fisher information matrix. In practice, as shown by Salimbeni et al. (2018), calculation of Fisher information matrix is not required as $\mathrm{F}(\boldsymbol{\eta}_t)^{-1} \nabla_{\boldsymbol{\eta}} \mathfrak{L}(\boldsymbol{\eta}_t) = \nabla_{\boldsymbol{\mu}} \mathfrak{L}(\boldsymbol{\mu}_t)$ where $\boldsymbol{\eta}$ is the natural parameter and $\boldsymbol{\mu}$ is the mean parameter.

Therefore, let, the natural parameters of $w(\mathbf{x}(\mathbf{z}))$ be $\boldsymbol{\eta}_w$ and the mean parameter be $\boldsymbol{\mu}_w$. In the following sections, the parameters without subscript is used for w distribution. As discussed in Raskutti & Mukherjee (2015), natural gradient descent can be cast as a mirror descent in mean parameterization

$$\boldsymbol{\mu}_{t+1} = \arg\min_{\boldsymbol{\mu}} \langle \nabla_{\boldsymbol{\mu}} \mathfrak{L}(\boldsymbol{\mu}_t), \boldsymbol{\mu} - \boldsymbol{\mu}_t \rangle + \frac{1}{\rho_t} \mathrm{D}_{\mathrm{KL}} \left[\boldsymbol{\mu} \,\|\, \boldsymbol{\mu}_t \right]. \tag{4.17}$$

Further, using the property $\partial_{\boldsymbol{\mu}} \mathrm{D}_{\mathrm{KL}} \left[\boldsymbol{\mu} \,\|\, \boldsymbol{\mu}_t \right] = \boldsymbol{\eta} - \boldsymbol{\eta}_t$

$$\boldsymbol{\mu}_{t+1} = \arg\min_{\boldsymbol{\mu}} \langle \nabla_{\boldsymbol{\mu}} \mathfrak{L}(\boldsymbol{\mu}_t), \boldsymbol{\mu} - \boldsymbol{\mu}_t \rangle + \frac{1}{\rho_t} \left(\boldsymbol{\eta} - \boldsymbol{\eta}_t \right). \tag{4.18}$$

As the derivative is zero at an extremum, for the ELBO in Eq. (4.13)

$$\begin{aligned}
&\nabla_{\boldsymbol{\mu}} \left[-\ell(\boldsymbol{\mu}_t) \right] + \frac{1}{\rho_t} \left(\boldsymbol{\eta} - \boldsymbol{\eta}_t \right) = 0 \\
&\nabla_{\boldsymbol{\mu}} \ell(\boldsymbol{\mu}_t) - \frac{1}{\rho_t} \left(\boldsymbol{\eta} - \boldsymbol{\eta}_t \right) = 0 \\
&\nabla_{\boldsymbol{\mu}} g - \partial_{\boldsymbol{\mu}} \mathrm{D}_{\mathrm{KL}} \left[w(\mathbf{x_z}) \,\|\, r(\mathbf{x_z}) \right] - \frac{1}{\rho_t} \left(\boldsymbol{\eta} - \boldsymbol{\eta}_t \right) = 0 \\
&\nabla_{\boldsymbol{\mu}} g - \boldsymbol{\eta} + \boldsymbol{\eta}_r - \frac{1}{\rho_t} \left(\boldsymbol{\eta} - \boldsymbol{\eta}_t \right) = 0,
\end{aligned} \tag{4.19}$$

where $g = \int_{t=0}^{\tau} \mathbb{E}_{q_\phi(\mathbf{x}_t)} \left[h_1(\mathbf{x}_t) \right] \, \mathrm{d}t + \sum_n \mathbb{E}_{q_\phi(\mathbf{x}_n)} \left[h_2(\mathbf{x}_n) \right]$.

Following Khan & Lin (2017), using the model definition Eq. (4.2) and the conjugate properties, it is known that $\boldsymbol{\eta} = \boldsymbol{\eta}_r + \bar{\boldsymbol{\lambda}}$. Therefore, Eq. (4.19) is evaluated as

$$\nabla_{\boldsymbol{\mu}} g - \bar{\boldsymbol{\lambda}} - \frac{1}{\rho_t} \left(\boldsymbol{\eta}_r + \bar{\boldsymbol{\lambda}} - \boldsymbol{\eta}_r - \bar{\boldsymbol{\lambda}}_t \right) = 0$$

$$\rho_t \nabla_{\boldsymbol{\mu}} g - \bar{\boldsymbol{\lambda}} (\rho_t + 1) + \bar{\boldsymbol{\lambda}}_t = 0. \tag{4.20}$$

Thus, the natural gradient update is given by

$$\begin{aligned} \bar{\boldsymbol{\lambda}}_{(t+1)} (\rho_t + 1) &= \rho_t \nabla_{\boldsymbol{\mu}} g + \bar{\boldsymbol{\lambda}}_t \\ \bar{\boldsymbol{\lambda}}_{(t+1)} &= r_t \nabla_{\boldsymbol{\mu}} g + (1 + r_t) \bar{\boldsymbol{\lambda}}_t, \end{aligned} \tag{4.21}$$

where $r_t = \frac{1}{1+\rho_t}$.

4.4 Natural gradient updates

For the natural gradient updates Eq. (4.21), the gradient of g term needs to be calculated. For it, first the variational posterior and the chain rule is derived.

Variational posterior

Following Adam et al. (2020, equation 13), using the state-space parameters, the conditional of sparse Markovian GP is

$$r(\mathbf{x}_t \mid \mathbf{x}_z) \sim \mathcal{N}(\mathbf{P}_t \, \mathbf{v}_t, \mathbf{T}_t), \tag{4.22}$$

where $\mathbf{v}_t = (\mathbf{u}_{t-}, \mathbf{u}_{t+})$ are the inducing variable pairs. Considering the probability density over the inducing variables to be Gaussian $w(\mathbf{x}_\mathbf{z}) \sim \mathcal{N}(\boldsymbol{\mu}_{w_\mathbf{z}}, \boldsymbol{\Sigma}_{w_\mathbf{z} \, w_\mathbf{z}})$, the variational posterior $q(\mathbf{x}_t)$ is written as

$$\begin{aligned} q(\mathbf{x}_t) &= r(\mathbf{x}_t \mid \mathbf{x}_z) \; w(\mathbf{x}_z), \\ q(\mathbf{x}_t) &\sim \mathcal{N}(\mathbf{P}_t \, \boldsymbol{\mu}_{w_t}, \mathbf{T}_t + \mathbf{P}_t \, \boldsymbol{\Sigma}_{w_t w_t} \, \mathbf{P}_t^\top) \\ &\sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \end{aligned} \tag{4.23}$$

where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{P}_t \in \mathbb{R}^{d \times 2d}$, $\mathbf{T}_t \in \mathbb{R}^{d \times d}$, $\mathbf{v}_t \in \mathbb{R}^{2d}$, $\boldsymbol{\mu}_{w_t} \in \mathbb{R}^{2d}$, $\boldsymbol{\Sigma}_{w_t w_t} \in \mathbb{R}^{2d \times 2d}$, $\boldsymbol{\mu}_t \in \mathbb{R}^d$, and $\boldsymbol{\Sigma}_t \in \mathbb{R}^{d \times d}$.

Chain rule

From Eq. (4.23), a chain rule to calculate the gradient with respect to $\boldsymbol{\mu}_{w_t}$ and $\boldsymbol{\Sigma}_{w_t w_t}$ is derived. For example, the gradient of $f_1(.)$ with respect to $\boldsymbol{\Sigma}_{w_t w_t}$ is calculated as

$$\nabla_{\boldsymbol{\Sigma}_{w_t w_t}} f_1(.) = \nabla_{\boldsymbol{\Sigma}_t} f_1(.) \cdot \nabla_{\boldsymbol{\Sigma}_{w_t w_t}} \boldsymbol{\Sigma}_t. \tag{4.24}$$

Therefore, following Eq. (4.23) and Eq. (4.24), the gradient of g with respect to $\boldsymbol{\mu}_{w_t}$ is

$$\partial_{\boldsymbol{\mu}_{w_t}} g = \int_\tau \mathbf{P}_\tau^\top \, \partial_{\boldsymbol{\mu}_\tau} \, g_1(\tau) \, \mathrm{d}\tau + \sum_n \mathbf{P}_n^\top \, \partial_{\boldsymbol{\mu}_n} \, g_2(n), \tag{4.25}$$

where $g_1(\tau) = \mathbb{E}_{q(\mathbf{x}_\tau)}\left[h_1(\mathbf{x}_\tau)\right]$ and $g_2(n) = \mathbb{E}_{q(\mathbf{x}_n)}\left[h_2(\mathbf{x}_n)\right]$. Similarly, the gradient with respect to $\boldsymbol{\Sigma}_{w_t w_t}$ is

$$\partial_{\boldsymbol{\Sigma}_{w_t w_t}} g = \int_\tau \mathbf{P}_\tau^\top \, \partial_{\boldsymbol{\Sigma}_\tau} \, g_1(\tau) \mathbf{P}_\tau \, \mathrm{d}\tau \, + \sum_n \mathbf{P}_n^\top \, \partial_{\boldsymbol{\Sigma}_n} \, g_2(n) \mathbf{P}_n. \tag{4.26}$$

For an exponential distribution, it is known that the mean parameter is $\mathbb{E}[\phi]$ where $\phi = [\mathbf{x}\,,\, \mathbf{x}^2]^\top$ is the sufficient statistics. Therefore,

$$\begin{aligned} &\mathbb{E}[\phi] = [\boldsymbol{\mu}^{(1)}\,,\, \boldsymbol{\mu}^{(2)}] = \left[\boldsymbol{\mu}_{w_t}\,,\, \boldsymbol{\mu}_{w_t}\boldsymbol{\mu}_{w_t}^\top + \boldsymbol{\Sigma}_{w_t w_t}\right]\,, \\ &\boldsymbol{\mu}_{w_t} = \boldsymbol{\mu}^{(1)}\,, \\ &\boldsymbol{\Sigma}_{w_t w_t} = \boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}\boldsymbol{\mu}^{(1)\top}. \end{aligned} \tag{4.27}$$

The gradients using the chain rule are calculated as

$$\begin{aligned} \partial_{\boldsymbol{\mu}^{(2)}} g &= \partial_{\boldsymbol{\Sigma}_{w_t w_t}} g \cdot \partial_{\boldsymbol{\mu}^{(2)}} \boldsymbol{\Sigma}_{w_t w_t} \\ &= \partial_{\boldsymbol{\Sigma}_{w_t w_t}} g, \end{aligned} \tag{4.28}$$

$$\begin{aligned} \partial_{\boldsymbol{\mu}^{(1)}} g &= \partial_{\boldsymbol{\Sigma}_{w_t w_t}} g \cdot \partial_{\boldsymbol{\mu}^{(1)}} \boldsymbol{\Sigma}_{w_t w_t} \\ &= \partial_{\boldsymbol{\Sigma}_{w_t w_t}} g \cdot (\partial_{\boldsymbol{\mu}^{(1)}} \, \boldsymbol{\mu}^{(2)} - 2\boldsymbol{\mu}^{(1)}) \\ &= \partial_{\boldsymbol{\Sigma}_{w_t w_t}} g \cdot \partial_{\boldsymbol{\mu}^{(1)}} \, \boldsymbol{\mu}^{(2)} - \partial_{\boldsymbol{\Sigma}_{w_t w_t}} g \times 2\boldsymbol{\mu}^{(1)} \\ &= \partial_{\boldsymbol{\mu}_{w_t}} g - \partial_{\boldsymbol{\Sigma}_{w_t w_t}} g \cdot 2\boldsymbol{\mu}_{w_t}. \end{aligned} \tag{4.29}$$

Therefore, by using Eq. (4.25) and Eq. (4.26) in Eq. (4.28) the gradient is calculated as

$$\partial_{\boldsymbol{\mu}^{(2)}} g = \int_\tau \mathbf{P}_\tau^\top \, \partial_{\boldsymbol{\Sigma}_\tau} \, g_1(\tau) \mathbf{P}_\tau \, \mathrm{d}\tau \, + \sum_n \mathbf{P}_n^\top \, \partial_{\boldsymbol{\Sigma}_n} \, g_2(n) \mathbf{P}_n, \tag{4.30}$$

$$\begin{aligned} \partial_{\boldsymbol{\mu}^{(1)}} g = &\int_\tau \mathbf{P}_\tau^\top \, \partial_{\boldsymbol{\mu}_\tau} \, g_1(\tau) \, \mathrm{d}\tau \, + \sum_n \mathbf{P}_n^\top \, \partial_{\boldsymbol{\mu}_n} \, g_2(n) \\ &- 2 \left(\int_\tau \mathbf{P}_\tau^\top \, \partial_{\boldsymbol{\Sigma}_\tau} \, g_1(\tau) \mathbf{P}_\tau \, \boldsymbol{\mu}_{w_\tau} \, \mathrm{d}\tau + \sum_n \mathbf{P}_n^\top \, \partial_{\boldsymbol{\Sigma}_n} \, g_2(n) \mathbf{P}_n \, \boldsymbol{\mu}_{w_n} \right). \end{aligned} \tag{4.31}$$

For a Gaussian distribution it is known that the natural parameters of the distribution are $\boldsymbol{\lambda} = \left[\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}\,,\, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\right]$ and $\partial_{\boldsymbol{\Sigma}_n} \, g_1(n) = \frac{1}{2}\partial^2_{\boldsymbol{\mu}_n \, \boldsymbol{\mu}_n} \, g_1(n)$. Thus, the gradients are evaluated as

$$\partial_{\boldsymbol{\mu}^{(2)}} g = \frac{1}{2} \left[\int_\tau \mathbf{P}_\tau^\top \, \partial^2_{\boldsymbol{\mu}_\tau \, \boldsymbol{\mu}_\tau} \, g_1(\tau) \mathbf{P}_\tau \, \mathrm{d}\tau \, + \sum_n \mathbf{P}_n^\top \, \partial^2_{\boldsymbol{\mu}_n \, \boldsymbol{\mu}_n} \, g_2(n) \mathbf{P}_n \right], \tag{4.32}$$

$$\begin{aligned} \partial_{\boldsymbol{\mu}^{(1)}} g = &\int_\tau \mathbf{P}_\tau^\top \, \partial_{\boldsymbol{\mu}_\tau} \, g_1(\tau) \, \mathrm{d}\tau \, + \sum_n \mathbf{P}_n^\top \, \partial_{\boldsymbol{\mu}_n} \, g_2(n) \\ &\int_\tau \mathbf{P}_\tau^\top \, \partial^2_{\boldsymbol{\mu}_\tau \, \boldsymbol{\mu}_\tau} \, g_1(\tau) \mathbf{P}_\tau \, \boldsymbol{\mu}_{w_\tau} \, \mathrm{d}\tau + \sum_n \mathbf{P}_n^\top \, \partial^2_{\boldsymbol{\mu}_n \, \boldsymbol{\mu}_n} \, g_2(n) \mathbf{P}_n \, \boldsymbol{\mu}_{w_n}. \end{aligned} \tag{4.33}$$

However, above gradient calculation requires Gaussian approximation which can be avoided for some distributions by using Price's and Bonnet's theorem.

Price's and Bonnet's theorem

Suppose, f is a scalar, non-linear function of $\{\mathbf{x}_i\}_{i=1}^n$ which are jointly distributed with $\boldsymbol{\mu}$ mean and $\boldsymbol{\Sigma}_{pg}$ covariance between $\mathbf{x}_p$, $\mathbf{x}_q$. Price's theorem (Price, 1958) states

$$\nabla_{\rho_{pq}} \mathbb{E}[f] = \mathbb{E}\left[\nabla^2_{\mathbf{x}_p\mathbf{x}_q} f\right], \tag{4.34}$$

where $p \neq q$ and the expectation is under a multi-variate Gaussian with all variances equal to one and the correlation coefficients ρ_{pq}. Bonnet's theorem (Bonnet, 1964) states

$$\nabla_{\boldsymbol{\mu}_t} \mathbb{E}[f] = \mathbb{E}\left[\nabla_{\mathbf{x}_t} f\right], \tag{4.35}$$

where $\mathbf{m}_t$ is the mean for $\mathbf{x}_t$.

Following the two theorems, the terms in Eq. (4.30) are evaluated as

$$\begin{aligned}
\textstyle\sum_n \mathbf{P}_n^\top \partial_{\boldsymbol{\Sigma}_n} g_2(n)\mathbf{P}_n &= \textstyle\sum_n \mathbf{P}_n^\top \partial_{\boldsymbol{\Sigma}_n} \mathbb{E}_{q(\mathbf{x}_n)}\left[h_2(\mathbf{x}_n)\right]\mathbf{P}_n \\
&= \textstyle\sum_n \mathbf{P}_n^\top \mathbb{E}_{q(\mathbf{x}_n)}\left[\nabla^2_{\mathbf{x}\mathbf{x}} h_2(\mathbf{x}_n)\right]\mathbf{P}_n\,, && (4.36)\\
\textstyle\sum_n \mathbf{P}_n^\top \partial_{\boldsymbol{\mu}_n} g_2(n) &= \textstyle\sum_n \mathbf{P}_n^\top \partial_{\boldsymbol{\mu}_n} \mathbb{E}_{q(\mathbf{x}_n)}\left[h_2(\mathbf{x}_n)\right] \\
&= \textstyle\sum_n \mathbf{P}_n^\top \mathbb{E}_{q(\mathbf{x}_n)}\left[\nabla_{\mathbf{x}} h_2(\mathbf{x}_n)\right]\,, && (4.37)\\
\textstyle\int_\tau \mathbf{P}_\tau^\top \partial_{\boldsymbol{\Sigma}_\tau} g_1(\tau)\mathbf{P}_\tau\,\mathrm{d}\tau &= \textstyle\int_\tau \mathbf{P}_\tau^\top \partial_{\boldsymbol{\Sigma}_\tau} \mathbb{E}_{q(\mathbf{x}_\tau)}\left[h_1(\mathbf{x}_\tau)\right]\mathbf{P}_\tau\,\mathrm{d}\tau \\
&= \textstyle\int_\tau \mathbf{P}_\tau^\top \mathbb{E}_{q(\mathbf{x}_\tau)}\left[\nabla^2_{\mathbf{x}\mathbf{x}} h_1(\mathbf{x}_\tau)\right]\mathbf{P}_\tau\,\mathrm{d}\tau\,, && (4.38)\\
\textstyle\int_\tau \mathbf{P}_\tau^\top \partial_{\boldsymbol{\mu}_\tau} g_1(\tau)\,\mathrm{d}\tau &= \textstyle\int_\tau \mathbf{P}_\tau^\top \partial_{\boldsymbol{\mu}_\tau} \mathbb{E}_{q(\mathbf{x}_\tau)}\left[h_1(\mathbf{x}_\tau)\right]\,\mathrm{d}\tau \\
&= \textstyle\int_\tau \mathbf{P}_\tau^\top \mathbb{E}_{q(\mathbf{x}_\tau)}\left[\nabla_{\mathbf{x}} h_1(\mathbf{x}_\tau)\right]\,\mathrm{d}\tau\,. && (4.39)
\end{aligned}$$

Thus, the gradient of g is evaluated as

$$\begin{aligned}
\partial_{\boldsymbol{\mu}^{(2)}} g &= \textstyle\int_\tau \mathbf{P}_\tau^\top \mathbb{E}_{q(\mathbf{x}_\tau)}\left[\nabla^2_{\mathbf{x}\mathbf{x}} h_1(\mathbf{x}_\tau)\right]\mathbf{P}_\tau\,\mathrm{d}\tau \\
&\quad + \textstyle\sum_n \mathbf{P}_n^\top \mathbb{E}_{q(\mathbf{x}_n)}\left[\nabla^2_{\mathbf{x}\mathbf{x}} h_2(\mathbf{x}_n)\right]\mathbf{P}_n, && (4.40)\\
\partial_{\boldsymbol{\mu}^{(1)}} g &= \textstyle\int_\tau \mathbf{P}_\tau^\top \mathbb{E}_{q(\mathbf{x}_\tau)}\left[\nabla_{\mathbf{x}} h_1(\mathbf{x}_\tau)\right]\,\mathrm{d}\tau + \sum_n \mathbf{P}_n^\top \mathbb{E}_{q(\mathbf{x}_n)}\left[\nabla_{\mathbf{x}} h_2(\mathbf{x}_n)\right] \\
&\quad - 2\left(\textstyle\int_\tau \mathbf{P}_\tau^\top \mathbb{E}_{q(\mathbf{x}_\tau)}\left[\nabla^2_{\mathbf{x}\mathbf{x}} h_1(\mathbf{x}_\tau)\right]\mathbf{P}_\tau\boldsymbol{\mu}_{w_\tau}\,\mathrm{d}\tau\right. \\
&\quad \left. + \textstyle\sum_n \mathbf{P}_n^\top \mathbb{E}_{q(\mathbf{x}_n)}\left[\nabla^2_{\mathbf{x}\mathbf{x}} h_2(\mathbf{x}_n)\right]\mathbf{P}_n\,\boldsymbol{\mu}_{w_n}\right). && (4.41)
\end{aligned}$$

4.5 Parameterization for $w(x(z))$ and parameter count

One of the ways to learn the variational parameter ξ is to keep the other parameters, ψ and θ, constant. Also, a plausible method to parameterize the

distribution w is by defining the first two moments, $\boldsymbol{\mu}_{w_\mathbf{z}}$ and $\boldsymbol{\Sigma}_{w_\mathbf{z} w_\mathbf{z}}$. At the optimum ELBO, as the aim is to maximize it,

$$\begin{aligned} \xi^* &= \arg\max_{\xi} \ell(\theta,\, \psi,\, \xi), \\ \nabla_\xi \ell_{|\xi=\xi^*|} &= 0. \end{aligned} \tag{4.42}$$

The derivative of $\mathbb{E}_{q_\phi(\mathbf{x}_t)}[w(\mathbf{x}_t)]$ with respect to $\boldsymbol{\Sigma}_{w_t w_t}$, is calculated using chain rule Eq. (4.24) as

$$\begin{aligned} \nabla_{\boldsymbol{\Sigma}_{w_t w_t}} \mathbb{E}_{q_\phi(\mathbf{x}_t)}[w(\mathbf{x}_t)] &= \nabla_{\boldsymbol{\Sigma}_t} \mathbb{E}_{q_\phi(\mathbf{x}_t)}[w(\mathbf{x}_t)] \cdot \nabla_{\boldsymbol{\Sigma}_{w_t w_t}} \boldsymbol{\Sigma}_t \\ &= \mathbf{P}_t^\top \nabla_{\boldsymbol{\Sigma}_t} \mathbb{E}_{q_\phi(\mathbf{x}_t)}[w(\mathbf{x}_t)]\, \mathbf{P}_t. \end{aligned} \tag{4.43}$$

Similarly, the derivative of KL in Eq. (4.14) with respect to $\boldsymbol{\Sigma}_{w_\mathbf{z} w_\mathbf{z}}$ is

$$\nabla_{\boldsymbol{\Sigma}_{w_\mathbf{z} w_\mathbf{z}}} \mathrm{D}_{\mathrm{KL}}[w(\mathbf{x}_\mathbf{z}) \,\|\, r(\mathbf{x}_\mathbf{z})] = \frac{1}{2}\left[-\boldsymbol{\Sigma}^{-1}_{w_\mathbf{z} w_\mathbf{z}} + \mathbf{K}^{-1}\right]. \tag{4.44}$$

Therefore, derivative of the ELBO in Eq. (4.13) with respect to $\boldsymbol{\Sigma}_{w_\mathbf{z} w_\mathbf{z}}$ is

$$\begin{aligned} \nabla_{\boldsymbol{\Sigma}_{w_\mathbf{z} w_\mathbf{z}}} \ell = \nabla_{\boldsymbol{\Sigma}_{w_\mathbf{z} w_\mathbf{z}}} \textstyle\int_\tau \mathbb{E}_{q_\phi(\mathbf{x}_\tau)}[h_1(\mathbf{x}_\tau)]\, \mathrm{d}\tau + \nabla_{\boldsymbol{\Sigma}_{w_\mathbf{z} w_\mathbf{z}}} \sum_n \mathbb{E}_{q_\phi(\mathbf{x}_n)}[h_2(\mathbf{x}_n)] \\ + \frac{1}{2}\left[\boldsymbol{\Sigma}^{-1}_{w_\mathbf{z} w_\mathbf{z}} - \mathbf{K}^{-1}\right], \end{aligned} \tag{4.45}$$

which is further evaluated as

$$\begin{aligned} \nabla_{\boldsymbol{\Sigma}_{w_\mathbf{z} w_\mathbf{z}}} \ell = \textstyle\int_\tau \mathbf{P}_\tau^\top \nabla_{\boldsymbol{\Sigma}_\tau} \mathbb{E}_{q_\phi(\mathbf{x}_\tau)}[h_1(\mathbf{x}_\tau)]\, \mathbf{P}_\tau\, \mathrm{d}\tau \\ + \textstyle\sum_n \mathbf{P}_n^\top \nabla_{\boldsymbol{\Sigma}_n} \mathbb{E}_{q_\phi(\mathbf{x}_n)}[h_2(\mathbf{x}_n)]\, \mathbf{P}_n \\ + \frac{1}{2}\left[\boldsymbol{\Sigma}^{-1}_{w_\mathbf{z} w_\mathbf{z}} - \mathbf{K}^{-1}\right]. \end{aligned} \tag{4.46}$$

The optimal value of $\boldsymbol{\Sigma}^*_{w_\mathbf{z} w_\mathbf{z}}$ is calculated by setting the derivative to zero

$$\begin{aligned} \nabla_{\boldsymbol{\Sigma}_{w_\mathbf{z} w_\mathbf{z}}} \ell &= 0\,, \\ \boldsymbol{\Sigma}^{*-1}_{w_\mathbf{z} w_\mathbf{z}} &= \mathbf{K}^{-1} - 2\textstyle\int_\tau \mathbf{P}_\tau^\top \nabla_{\boldsymbol{\Sigma}_\tau} \mathbb{E}_{q_\phi(\mathbf{x}_\tau)}[h_1(\mathbf{x}_\tau)]\, \mathbf{P}_\tau\, \mathrm{d}\tau \\ &\quad - 2\textstyle\sum_n \mathbf{P}_n^\top \nabla_{\boldsymbol{\Sigma}_n} \mathbb{E}_{q_\phi(\mathbf{x}_n)}[h_2(\mathbf{x}_n)]\, \mathbf{P}_n \\ \boldsymbol{\Sigma}^{*-1}_{w_\mathbf{z} w_\mathbf{z}} &= \mathbf{K}^{-1} - \textstyle\int_\tau \boldsymbol{\alpha}_\tau \mathbf{P}_\tau \mathbf{P}_\tau^\top\, \mathrm{d}\tau - \sum_n \boldsymbol{\beta}_n \mathbf{P}_n\, \mathbf{P}_n^\top, \end{aligned} \tag{4.47}$$

where $\boldsymbol{\alpha}_\tau = 2\nabla_{\boldsymbol{\Sigma}_\tau} \mathbb{E}_{q_\phi(\mathbf{x}_\tau)}[h_1(\mathbf{x}_\tau)]$ and $\boldsymbol{\beta}_n = 2\nabla_{\boldsymbol{\Sigma}_n} \mathbb{E}_{q_\phi(\mathbf{x}_n)}[h_2(\mathbf{x}_n)]$.

Therefore, after discretising the integral term into t grid size, the total number of parameters required are total $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters that are $t + n$ parameters.

Chapter 5

Experiments

This chapter presents the experiment with Ornstein–Uhlenbeck (OU) process and a double-well process. For the OU process, the exact posterior is known thus this experiment is for a sanity check. A double-well process is a good candidate because of its non-linear nature. The primary aim of these experiments is to learn the posterior of the underlying SDE based on observation points over time. For both the processes, Gaussian process regression (GPR) using GPFlow (Matthews et al., 2017), GP-SDE using Numpy (Harris et al., 2020) and JAX (Bradbury et al., 2018), and the proposed method, SGP-SDE using Newt (Wilkinson, 2021) are performed. In addition to them, for the OU process, Doob's h-transform is also performed.

5.1 Ornstein–Uhlenbeck (OU) process

Ornstein–Uhlenbeck (OU) process is a stochastic process of a particle going through a Brownian motion (Uhlenbeck & Ornstein, 1930). It is a stationary Markovian GP, discussed in Section 2.4, and can be expressed by an SDE

$$\mathrm{d}\mathbf{x}(t) = -\boldsymbol{\lambda}\,\mathbf{x}(t)\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{\beta}(t), \tag{5.1}$$

where drift function is $f(\mathbf{x}_t) = -\boldsymbol{\lambda}\,\mathbf{x}_t$, diffusion function is $\boldsymbol{\sigma}$ and Brownian motion has $\mathbf{Q}$ spectral density.

OU process can also be thought of as a modification of a random walk which is mean-reverting, over time it drops to its mean, which can be observed in Figure 5.1.

Setup

An OU process trajectory using Euler–Maruyama is simulated, and 20 data points are randomly selected for the experiment, shown in Figure 5.2. The

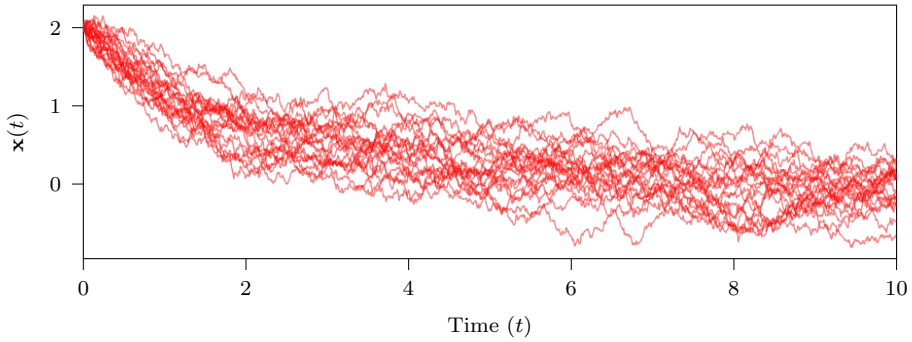

Figure 5.1: OU process trajectories simulated using Euler–Maruyama showcasing the mean-reverting nature of the process.

observation model is a Gaussian with a zero mean and 0.01 variance.

The parameters of the SDE used for simulating the trajectory are drift coefficient $\boldsymbol{\lambda} = 0.5$, diffusion coefficient $\boldsymbol{\sigma} = 1$, and spectral density of the Brownian motion $\mathbf{Q} = 0.1$. The initial state $\mathbf{x}_0 = 2$ and Euler–Maruyama is performed from time 0 to 10 with time-step 0.01. For all the methods performed, the same experimental setup with the same observation points is used.

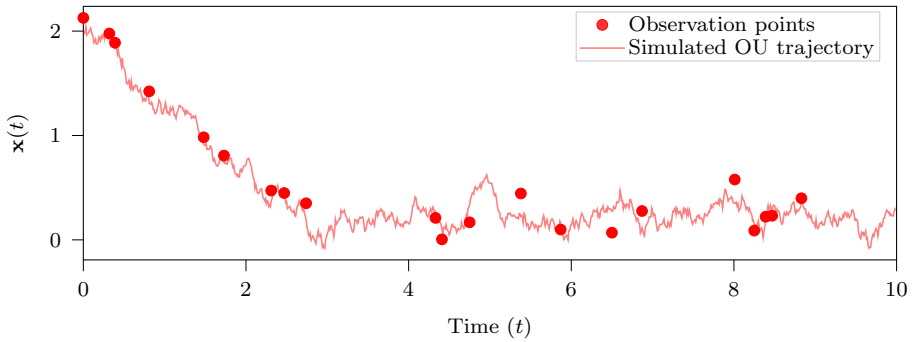

Figure 5.2: A OU process trajectory with randomly selected noisy observation points showcasing the experimental setup.

Gaussian process regression (GPR)

For the OU process, the kernel covariance function which provides the exact solution of the posterior is known. As shown by Särkkä & Solin (2019, Example 6.8), the stationary kernel covariance function corresponding to the

OU process is

$$\kappa(t,\, t') = \frac{\boldsymbol{\gamma}}{2\,\boldsymbol{\lambda}} \exp\left(-\boldsymbol{\lambda}\,|t - t'|\right), \tag{5.2}$$

where $\boldsymbol{\gamma} = \boldsymbol{\sigma}^2\,\mathbf{Q}$.
On conditioning a GP with OU kernel on the observed points, the exact posterior is calculated in closed form, shown in Figure 5.3.

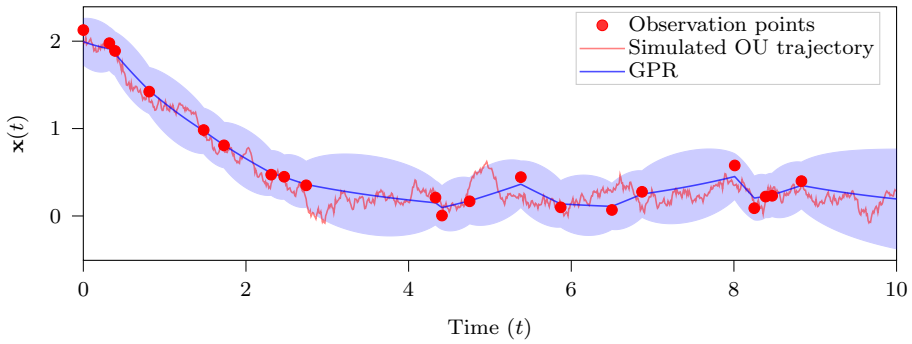

Figure 5.3: Mean and 95% confidence region of the posterior obtained for the OU process with noisy observation points by conditioning a GPR with OU kernel.

Doob's h-transform

Doob's h-transform is a method used to get an SDE by conditioning another SDE on its end point (Särkkä & Solin, 2019, Chapter 7). The primary idea of Doob's h-transform is that the result of multiplying the transition density of the original SDE with a term is an SDE which gives the transformed transition density.

As shown by the authors, the conditioned OU process can be written as

$$\mathrm{d}\mathbf{x} = \left[-\boldsymbol{\lambda}\,\mathbf{x} + \frac{\boldsymbol{\alpha}\,\mathbf{a}(t)}{\boldsymbol{\sigma}^2(t)}\left(\mathbf{x}_T - \mathbf{a}(t)\,\mathbf{x}\right)\right]\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}, \tag{5.3}$$

where

$$\begin{aligned}
\mathbf{a}(t) &= \exp\left(-\boldsymbol{\lambda}\,(T - t)\right), \\
\boldsymbol{\sigma}^2(t) &= \frac{\boldsymbol{\alpha}}{2\boldsymbol{\lambda}}\left[1 - \exp\left(-2\boldsymbol{\lambda}\,(T - t)\right)\right], \\
\boldsymbol{\alpha} &= q\,\boldsymbol{\sigma}^2.
\end{aligned}$$

Figure 5.4 showcases the first two moments of the Gaussian states of the conditioned OU SDE. From the figure, it can also be inferred that the conditioned GP is identical to the posterior of the GPR with OU kernel which is expected as both of them give the exact posterior.

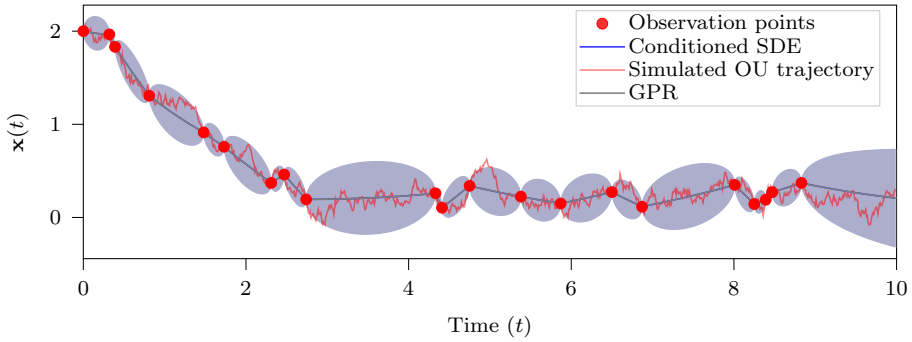

Figure 5.4: Mean and 95% confidence interval of the posterior obtained for the OU process with noise-less observation points obtained by GPR with OU kernel and Doob's h-transform (conditioned SDE). Both the posteriors are identical, expressing that both methods are fundamentally the same.

GP-SDE

As discussed in Section 3.2, the SDE can be approximated path-wise via Gaussian process approximation as

$$\begin{aligned} \mathrm{d}\mathbf{x}(t) &= f(\mathbf{x}(t))\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{\beta}(t) \\ &\approx f_L(\mathbf{x}(t))\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{\beta}(t), \end{aligned} \tag{5.4}$$

where $f_L(\mathbf{x}(t)) = -\mathbf{A}(t)\,\mathbf{x}(t) + \mathbf{b}(t)$.

Following the discussion, variational approximation is used with the ELBO as the objective along with the conditions incorporated by performing constrained optimization. For the OU process, the E_{sde} in Eq. (3.15) is simplified to

$$E_{sde}(t) = \frac{1}{2\,\boldsymbol{\sigma}^2}\left[(\mathbf{A}_t - \boldsymbol{\alpha})^2\left\langle \mathbf{x}_t^2 \right\rangle_{q(\mathbf{x}_t)} - 2\,\mathbf{b}_t\,(\mathbf{A}_t - \boldsymbol{\alpha})\left\langle \mathbf{x}_t \right\rangle_{q(\mathbf{x}_t)} + \mathbf{b}_t^2\right], \tag{5.5}$$

where $\boldsymbol{\alpha}$ is the drift coefficient. Also, the update rules Eq. (3.23) can be written as

$$\bar{\mathbf{A}}_t = \boldsymbol{\alpha} + 2\boldsymbol{\sigma}^2\boldsymbol{\Psi}_t, \tag{5.6}$$

$$\bar{\mathbf{b}}_t = (\bar{\mathbf{A}}_t - \boldsymbol{\alpha})\mathbf{m}_t - \boldsymbol{\sigma}^2\boldsymbol{\lambda}_t, \tag{5.7}$$

where $\boldsymbol{\lambda}_t$ and $\boldsymbol{\Psi}_t$ are the Lagrange multipliers.

Figure 5.5 showcases the approximated posterior of the OU process along with the evolution of parameters of $f_L(\cdot)$ and the Lagrange multipliers over time. These parameters jump when a value is observed due to the jump conditions.

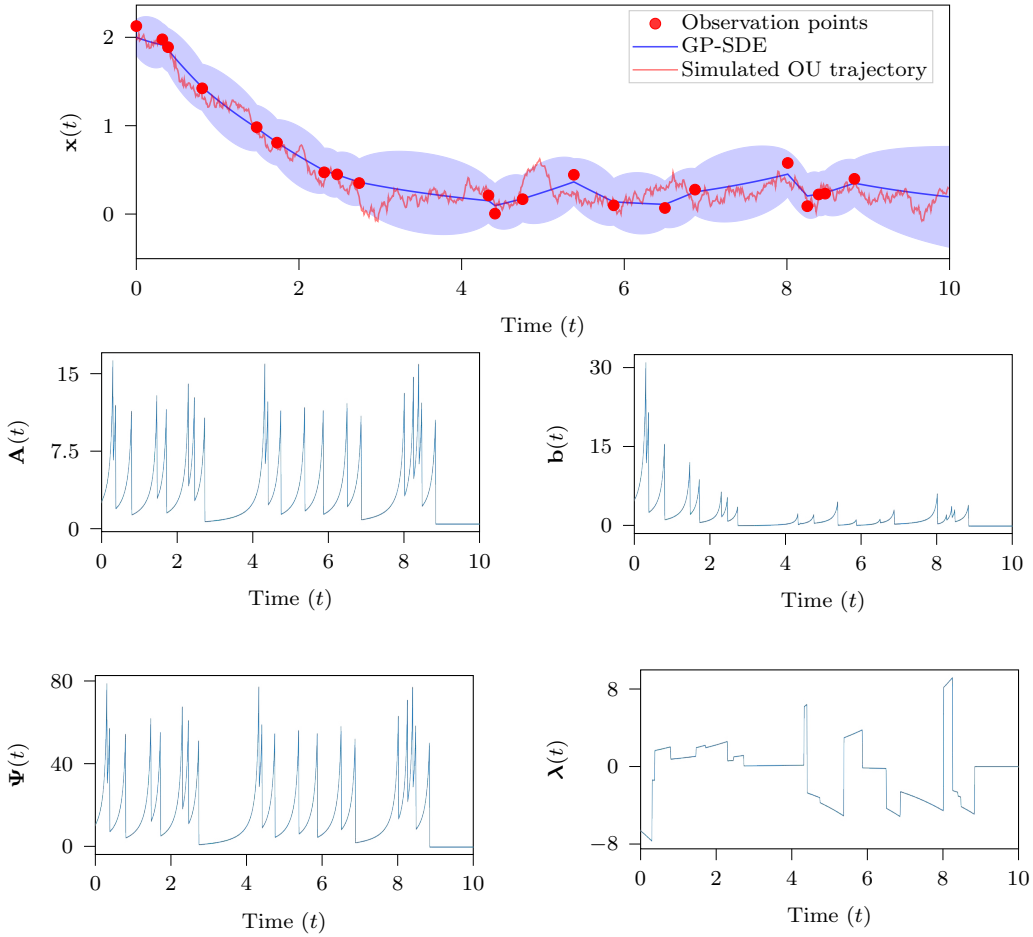

Figure 5.5: Mean and 95% confidence interval of the posterior obtained by GP-SDE for the OU process with noisy observation points and the Lagrange multiplier values, $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\Psi}(t)$, and the variational parameters $\mathbf{A}(t)$ and $\mathbf{b}(t)$.

SGP-SDE

The proposed method, SGP-SDE, is employed in the same experimental setup to approximate the posterior. Thus,

$$p_\theta(\mathbf{x}(\cdot)) : \ \mathrm{d}\mathbf{x}(t) = -\boldsymbol{\lambda}(\mathbf{x}(\cdot))\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{\beta}(t), \tag{5.8}$$

$$q_\phi(\mathbf{x}(\cdot)) : \ r(\mathbf{x}(\cdot) \mid \mathbf{x}(\mathbf{z}))\ w(\mathbf{x}(\mathbf{z})), \tag{5.9}$$

where for r, Matérn 1/2 kernel is chosen. The inducing variables are taken to be the same as the observation points and are not optimized.

For OU process, Girsanov's term in the ELBO Eq. (4.13) is simplified to

$$\ell_{\text{Girsanov}} = -\frac{1}{2\,\boldsymbol{\sigma}^2} \int_\tau (\boldsymbol{\lambda} + F)^2 (\mathbf{m}_\tau^2 + \mathbf{S}_\tau)\, \mathrm{d}\tau, \tag{5.10}$$

where the integral can be calculated as Riemann sums. Similarly, the natural gradient updates Eq. (4.32) for the Girsanov's term can be calculated in closed form as

$$g_1(\tau) = -\frac{1}{2\,\boldsymbol{\sigma}^2}(\boldsymbol{\lambda} + F)^2(\boldsymbol{\mu}_\tau^2 + \boldsymbol{\Sigma}_\tau), \tag{5.11}$$

$$\partial_{\boldsymbol{\mu}_\tau}\, g_1(\tau) = -\frac{1}{\boldsymbol{\sigma}^2}(\boldsymbol{\lambda} + F)^2 \boldsymbol{\mu}_\tau, \tag{5.12}$$

$$\partial^2_{\boldsymbol{\mu}_\tau\, \boldsymbol{\mu}_\tau}\, g_1(\tau) = -\frac{1}{\boldsymbol{\sigma}^2}(\boldsymbol{\lambda} + F)^2. \tag{5.13}$$

The posterior obtained by SGP-SDE is showcased in Figure 5.6 and the ELBO values along with its terms are plotted in Figure 5.7.

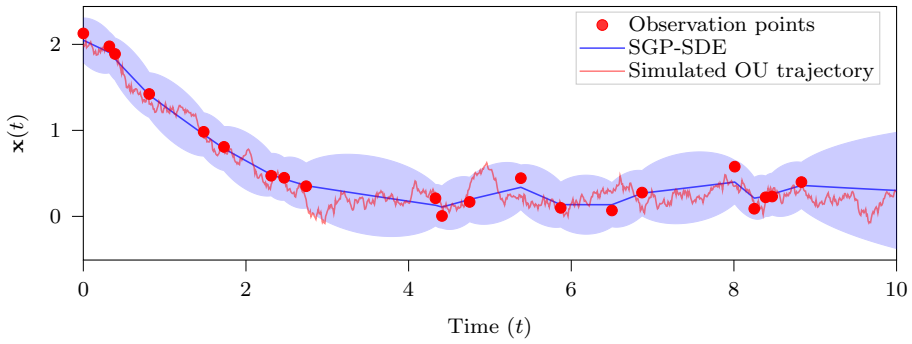

Figure 5.6: Mean and 95% confidence interval of the posterior obtained by SGP-SDE for the OU process with noisy observation points.

5.2 Double-well experiment

Double-well system is a non-linear system which makes it an ideal candidate for the experiment. The drift of the double-well SDE arises from the potential

$$u(\mathbf{x}) = -2\,\mathbf{x}^2 + \mathbf{x}^4, \tag{5.14}$$

which leads to the following drift function

$$f(\mathbf{x}) = -\frac{\mathrm{d}u(\mathbf{x})}{\mathrm{d}\mathbf{x}} = 4\,\mathbf{x} - 4\,\mathbf{x}^3. \tag{5.15}$$

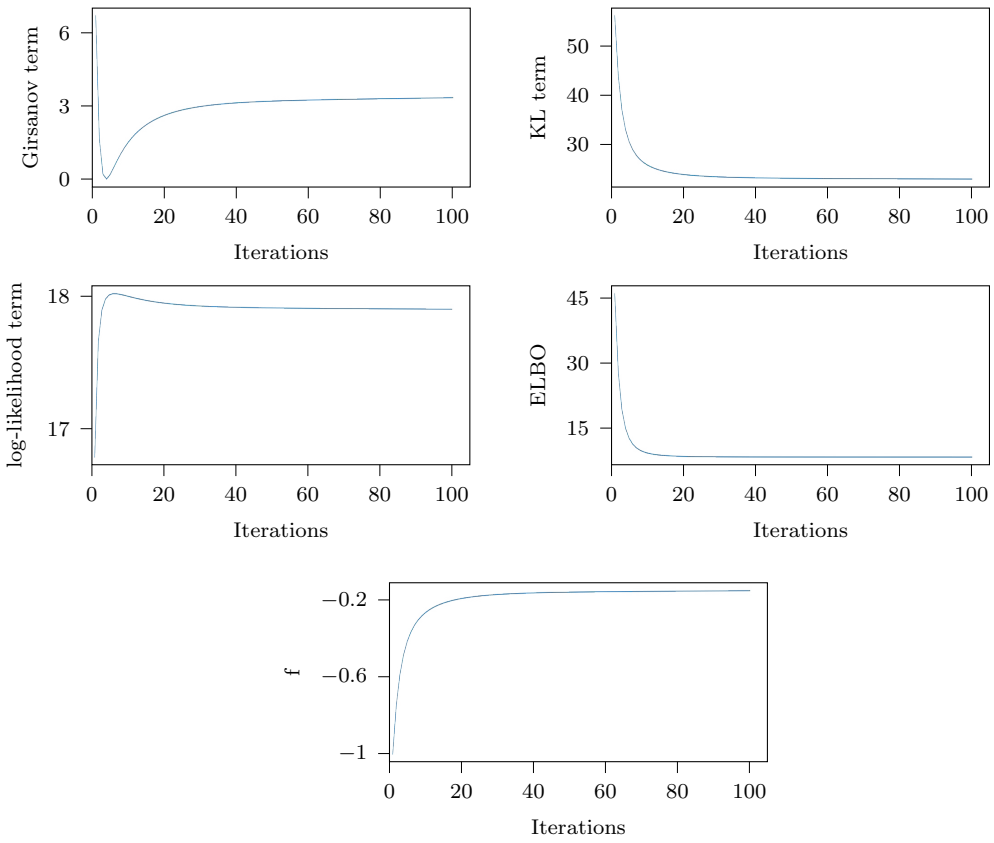

Figure 5.7: The evolution of the SGP-SDE method's ELBO terms and the drift function of the sparse Markovian GP r over iterations for the OU process with noisy observation points.

The double-well process has two minima $+1$ and -1 which is evident from the drift function $f(\mathbf{x})$ as well as from the sample Euler–Maruyama trajectory Figure 5.8. The state $\mathbf{x}(t)$ fluctuates between the two minima due to the driving noise which makes the process non-Gaussian.

Setup

A double-well process is simulated using Euler–Maruyama, and uniformly distributed data points at an interval of 0.5 are selected for the experiment, shown in Figure 5.9. The observation model is a Gaussian with a zero mean and 0.05 variance.

The parameters of the SDE used for simulating the trajectory are Eq. (5.15) as drift function, diffusion coefficient $\mathbf{L} = 1.22$, and spectral density of the Brownian motion $\mathbf{Q} = 0.5$. The initial state $\mathbf{x}_0 = 1$ and Euler–

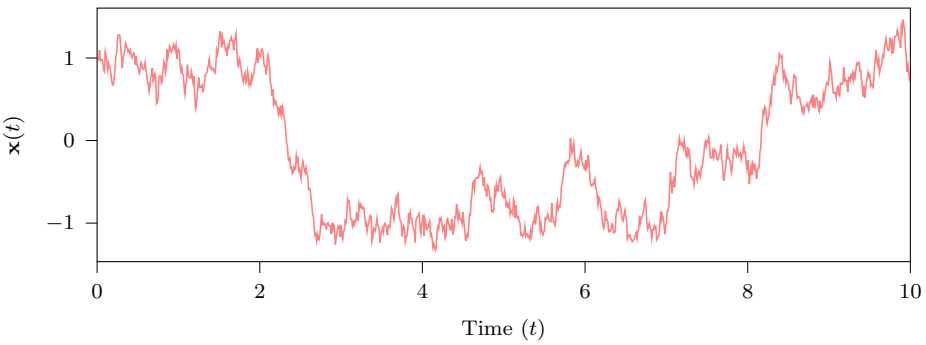

Figure 5.8: A double-well process trajectory simulated using Euler–Maruyama showcasing the two minima at +1 and −1 and the state fluctuations between them.

Maruyama is performed from time 0 to 10 with time-step 0.01. For all the methods performed, the same experimental setup with the same observation points is used.

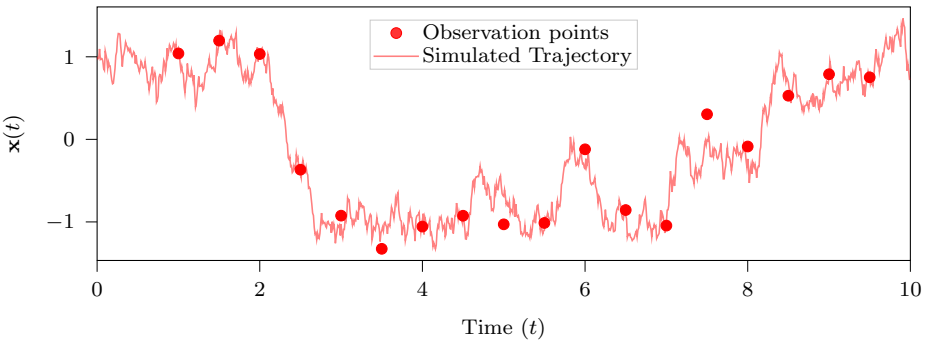

Figure 5.9: A double-well process trajectory with uniformly distributed noisy observation points showcasing the experimental setup.

Gaussian process regression (GPR)

A GP with squared-exponential kernel (RBF) is conditioned on the observed data points in order to obtain the posterior, showcased in Figure 5.10.

The observation model is fixed for the experiment and Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 0.01 is used to optimize the parameters with the aim to minimize the negative log-likelihood.

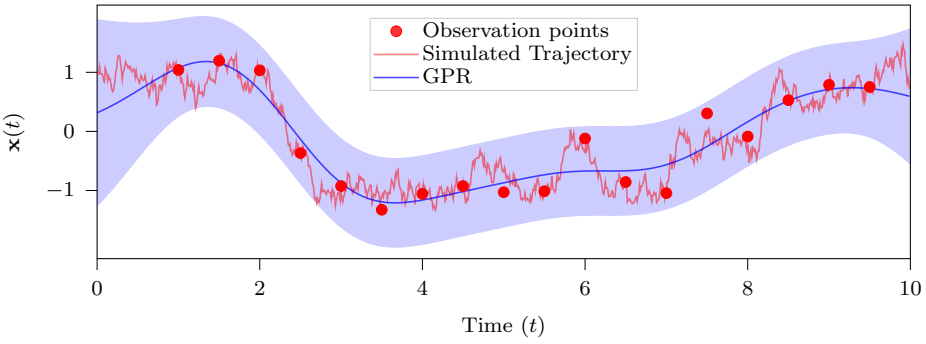

Figure 5.10: Mean and 95% confidence region of the posterior obtained for the double-well process with noisy observation points by conditioning a GPR with an RBF kernel.

GP-SDE

As discussed in Section 3.2 and similar to the OU experiment, the SDE can be approximated path-wise via Gaussian process approximation as

$$\begin{aligned} \mathrm{d}\mathbf{x}(t) &= f(\mathbf{x}(t))\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{\beta}(t) \\ &\approx f_L(\mathbf{x}(t))\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{\beta}(t), \end{aligned} \tag{5.16}$$

where $f_L(\mathbf{x}(t)) = -\mathbf{A}(t)\,\mathbf{x}(t) + \mathbf{b}(t)$. However, in this case, the drift of the prior SDE is non-linear so $f_L(\mathbf{x}(t))$ tries to approximate and learn the linearized drift function.

Following the discussion, variational approximation is used with the ELBO as the objective along with the conditions incorporated by performing constrained optimization. For the double-well process, the E_{sde} in Eq. (3.15) is written in closed form as

$$\begin{aligned} E_{sde}(t) = \frac{1}{2\,\boldsymbol{\sigma}^2} \big\langle 16\,\mathbf{x}_t^6 - 8(4+\mathbf{A}_t)\,\mathbf{x}_t^4 + 8\,\mathbf{b}_t\,\mathbf{x}_t^3 + (4+\mathbf{A}_t)^2\,\mathbf{x}_t^2 \\ + \mathbf{b}_t^2 - 2(4+\mathbf{A}_t)\,\mathbf{b}_t\,\mathbf{x}_t \big\rangle_{q(\mathbf{x}_t)} \,. \end{aligned} \tag{5.17}$$

Also, the update rules Eq. (3.23) is written as

$$\bar{\mathbf{A}}_t = -4(1 - 3\,\mathbf{m}_t^2 - 3\,\mathbf{S}_t) + 2\boldsymbol{\Psi}_t\,\boldsymbol{\sigma}^2, \tag{5.18}$$

$$\bar{\mathbf{b}}_t = -4\left\langle \mathbf{x}_t^3 \right\rangle_{q(\mathbf{x}_t)} + (4+\mathbf{A}_t)\mathbf{m}_t - \boldsymbol{\sigma}^2\boldsymbol{\lambda}_t, \tag{5.19}$$

where $\boldsymbol{\lambda}_t$ and $\boldsymbol{\Psi}_t$ are the lagrange multipliers and $q(\mathbf{x}_t) \sim \mathcal{N}(\mathbf{m}_t, \mathbf{S}_t)$.

Figure 5.11 showcases the approximated posterior of the double-well process along with the evolution of parameters of $f_L(\cdot)$ and the lagrange multipliers over time. These parameters jump when a value is observed due to the jump conditions.

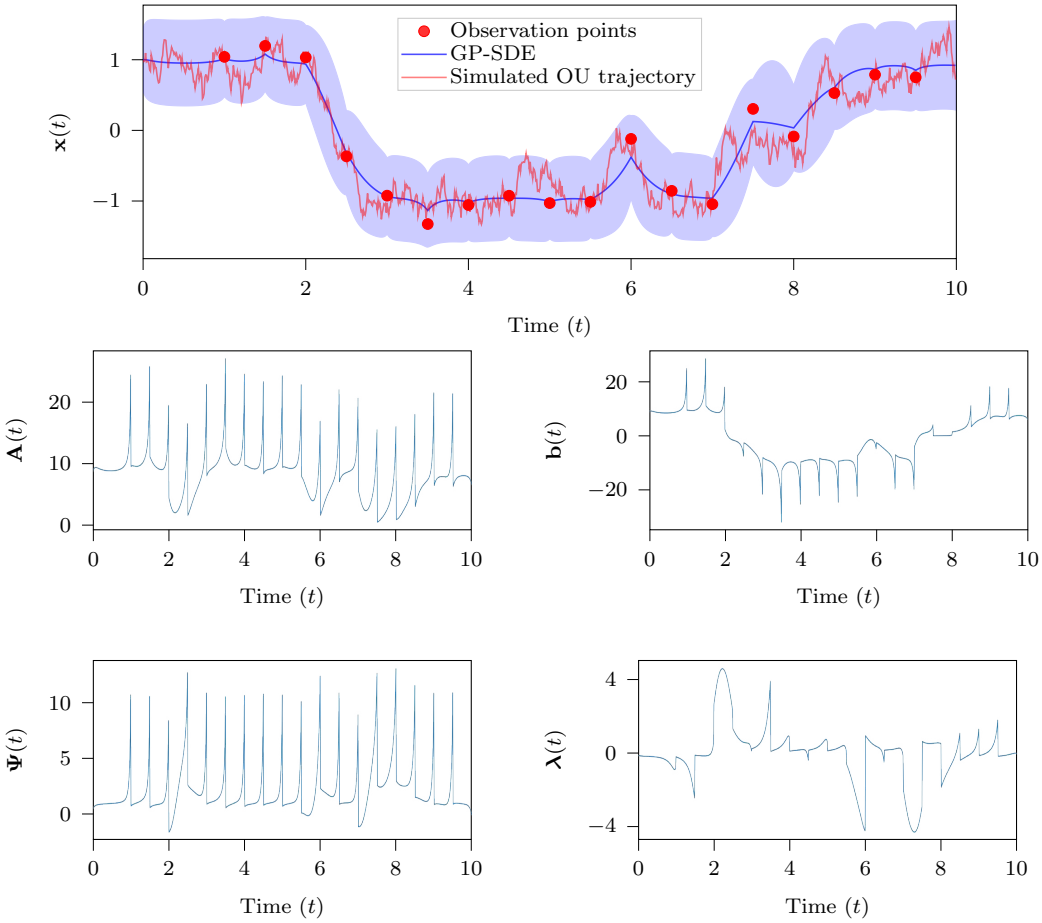

Figure 5.11: Mean and 95% confidence interval of the posterior obtained by GP-SDE for the double-well process with noisy observation points and the Lagrange multiplier values, $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\Psi}(t)$, and the variational parameters $\mathbf{A}(t)$ and $\mathbf{b}(t)$.

SGP-SDE

The proposed method, SGP-SDE, is employed in the same experimental setup to approximate the posterior. Thus, similar to the OU process experiment,

$$p_\theta(\mathbf{x}(\cdot)) : \ \mathrm{d}\mathbf{x}(t) = f(\mathbf{x}(\cdot))\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{\beta}(t), \tag{5.20}$$

$$q_\phi(\mathbf{x}(\cdot)) : \ r(\mathbf{x}(\cdot) \mid \mathbf{x}(\mathbf{z}))\ w(\mathbf{x}(\mathbf{z})), \tag{5.21}$$

where $f(\mathbf{x}) = 4\,\mathbf{x}(1-\mathbf{x}^2)$ and for r, Matérn 1/2 kernel is chosen. The inducing variables are taken to be the same as the observation points and are not

optimized.

For the double-well process, the expectation in the Girsanov's term in the ELBO Eq. (4.13) is approximated using Gaussian quadrature method and integral is calculated via Riemann sums.

The posterior obtained by SGP-SDE for double-well is showcased in Figure 5.12 and ELBO values along with its terms are plotted in Figure 5.13.

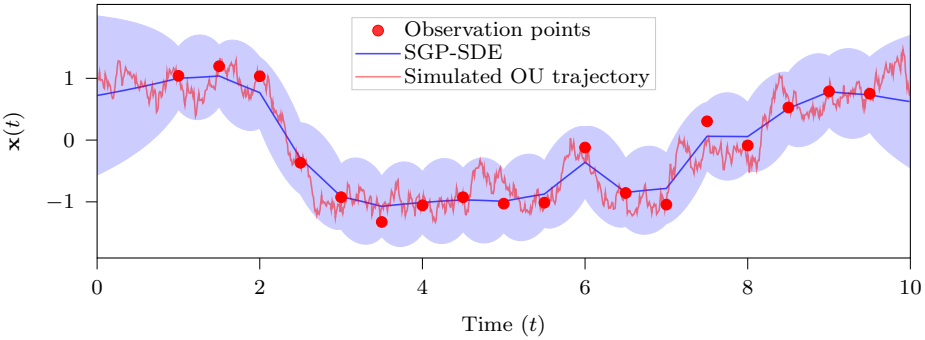

Figure 5.12: Mean and 95% confidence interval of the posterior obtained by SGP-SDE for the double-well process with noisy observation points

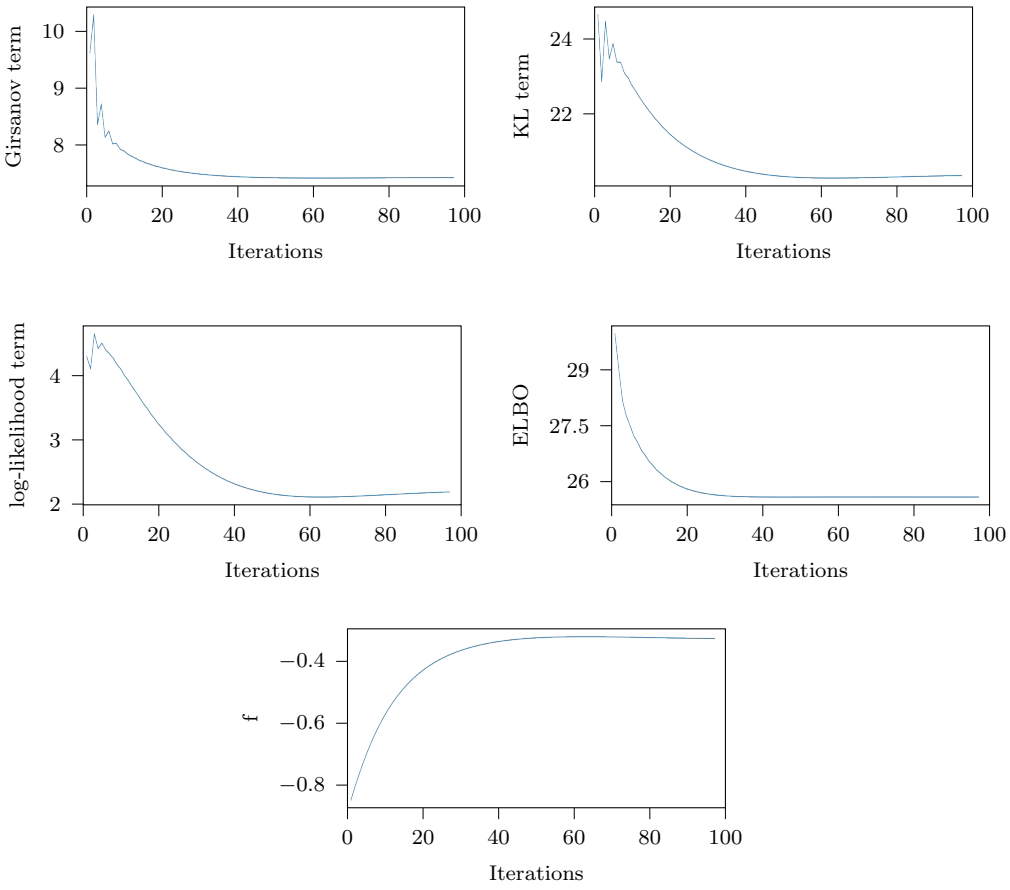

Figure 5.13: The evolution of the SGP-SDE method's ELBO terms and the drift function of the sparse Markovian GP r over iterations for the double-well process with noisy observation points.

Chapter 6

Discussion

The chapter discusses the two experiments, the Ornstein–Uhlenbeck (OU) process and the double-well process, and analysis the approximated posterior showcasing the inference capability of the proposed method. Both the experiments are discussed separately, followed by a discussion on the extension of the proposed method.

6.1 Ornstein–Uhlenbeck (OU) process

For the OU process, the exact posterior is obtained by Gaussian process regression (GPR) using the OU kernel and Doob's h-transform. Along with SGP-SDE, the GP-SDE method is also performed to approximate the posterior.

As discussed, SDE for the OU process is

$$\mathrm{d}\mathbf{x}(t) = -\boldsymbol{\lambda}\,\mathbf{x}(t)\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{\beta}(t)\,, \tag{6.1}$$

and for the experiment, the value of $\boldsymbol{\lambda} = 0.5$, $\boldsymbol{\sigma} = 1$, and $\mathbf{Q} = 0.1$. A linear SDE defines the process and, as the posterior can be evaluated in closed form, the experiment is primarily performed for a sanity check.

As shown in Figure 6.1, the posterior obtained by GP-SDE matches the exact posterior obtained by GPR. However, a deviation is noticed for the posterior obtained by SGP-SDE.

The Girsanov's term is expected to converge at zero but it does not as shown in Figure 5.7. Consequently, the lengthscale does not converge to the true lengthscale value, obtained on comparison of the OU kernel and the Matérn 1/2 kernel.

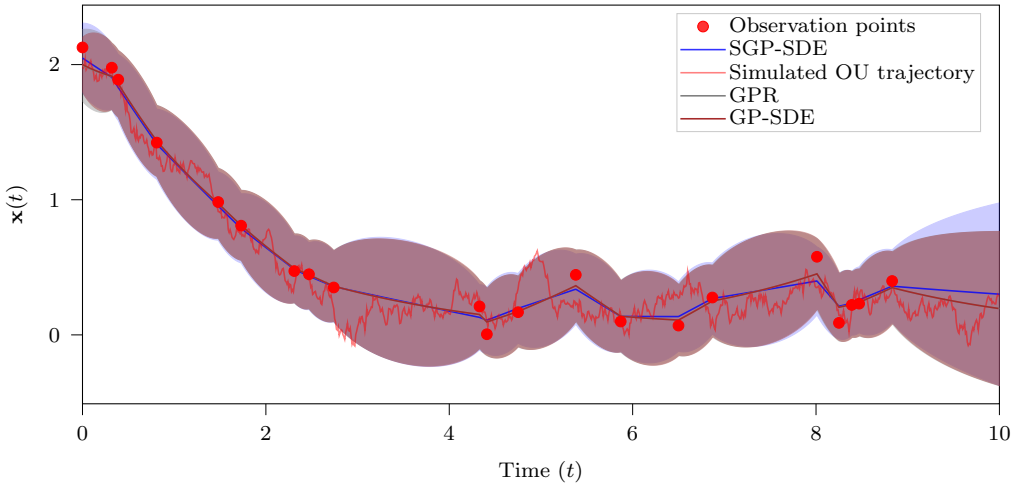

Figure 6.1: Mean and 95% confidence interval of the posterior obtained by GPR with OU kernel, GP-SDE, and SGP-SDE for the OU process with noisy observation points.

6.2 Double-well process

The SDE for the double-well process is

$$\mathrm{d}\mathbf{x}(t) = f(\mathbf{x}(t))\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{\beta}(t), \tag{6.2}$$

where $f(\mathbf{x}(t)) = 4\,\mathbf{x}(t)(1-\mathbf{x}(t)^2)$. To approximate the posterior, Gaussian process regression (GPR), GP-SDE, and SGP-SDE are performed.

The posterior obtained by GPR with an RBF kernel, shown in Figure 5.10, is not ideal as it does not identify the two wells. It is mainly because the GPR does not use the knowledge of the dynamics. Also, it requires high variance to explain the data of the two wells.

Figure 5.12 showcases the posterior approximated by the proposed method, SGP-SDE. As stationary Markovian GP is used, it fails to express the double-well process data which fluctuates between the two wells leading to a non-zero centered dynamics.

6.3 Limitation and extension

One of the benefits of the proposed method is that the sparse posterior is cheap in computation and easy to evaluate. However, a stationary GP, that is with a drift $f(\mathbf{x}(t), t) = \mathbf{F}\,\mathbf{x}(t)$, is not able to approximate SDEs like

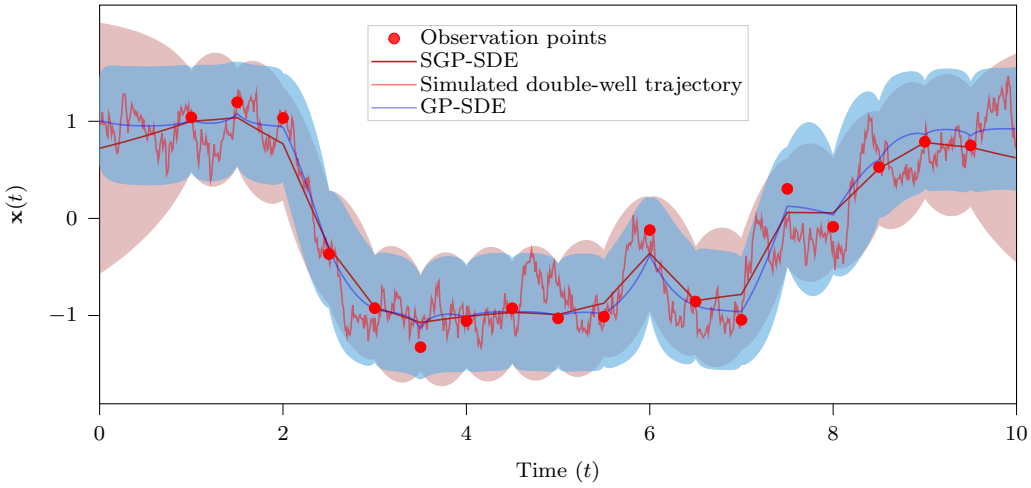

Figure 6.2: Mean and 95% confidence interval of the posterior obtained by GP-SDE and SGP-SDE for the double-well process with noisy observation points.

double-well globally. To overcome this, following extensions to the current method are proposed.

The current implementation of the proposed method involves the LTI SDE representation of the sparse markovian GP

$$\mathrm{d}\mathbf{x}(t) = \mathbf{F}\,\mathbf{x}(t)\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{\beta}(t). \tag{6.3}$$

However, by representing the Markovian GP as

$$\mathrm{d}\mathbf{x}(t) = \mathbf{F}\,(\mathbf{x}(t) - \mathbf{u})\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{\beta}(t)\,, \tag{6.4}$$

it is made more expressive as it allows to capture the dynamics of a non-zero mean-reverting process. This representation would benefit the double-well experiment as it has three specific gradient regions which can be incorporated using this representation. A piece-wise stationary kernel is also a plausible way to incorporate the information of a region-varying dynamics.

Chapter 7

Conclusion

The thesis proposes a novel method, SGP-SDE, to learn the stochastic differential equation (SDE) describing a dynamical system based on a set of discrete observations. The stochasticity in the dynamical system, discrete observations for continuous paths, complex underlying SDE and a complex observation model, make the task challenging.

Bayesian methods have been a popular choice for such tasks with an objective to maximize the marginal log-likelihood of the observations. However, it is intractable for most of the systems, and thus approximate algorithms are employed. Gaussian processes (GPs) are often used as approximate posterior over SDE paths. The resulting algorithms require fine discretization of the time horizon leading to a significant number of parameters and high complexity in both space and time.

By exploring the recent advances in approximate inference related to sparse GPs, the thesis presents an alternative parameterization to the approximate distribution over SDE paths based on a sparse Markovian Gaussian process. Similar to the current methods, the lower bound to the log-marginal likelihood, evidence lower bound (ELBO), is optimized. However, in contrast to the current methods, the proposed parameterization results in easy to evaluate, parallelizable ELBO. The resulting algorithm requires fewer number of parameters and reduces complexity, allowing well-defined optimization algorithms such as natural gradient descent for better convergence.

The method is evaluated on two processes: the Ornstein–Uhlenbeck (OU) process and the double-well process. Both processes are quite different from each other as the OU process is represented by a linear SDE whose true posterior is available in closed form. In contrast, a non-linear SDE represents the double-well process. Both the experiments demonstrate the capability of the proposed method to approximate the states posterior with a significantly

less number of parameters than the current methods.

The work presented in the thesis is preliminary. Future work involves making the sparse Markovian Gaussian posterior distribution more expressive and adopting a quantitative metric to evaluate the posterior between different methods. Furthermore, an experiment on a complex, real-world SDE with a multidimensional state vector is required to showcase the true capability of the method.

Bibliography

Adam, V., Eleftheriadis, S., Artemev, A., Durrande, N., and Hensman, J. Doubly sparse variational gaussian processes. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. Gaussian process approximations of stochastic differential equations. In *Gaussian Processes in Practice*. PMLR, 2007.

Archambeau, C., Opper, M., Shen, Y., Cornford, D., and Shawe-taylor, J. Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

Bertsekas, D. P. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1996.

Bhardwaj, R., Nambiar, A. R., and Dutta, D. A study of machine learning in healthcare. In *IEEE Forty-First Annual Computer Software and Applications Conference (COMPSAC)*, pp. 236–241, 2017.

Bishop, C. *Pattern Recognition and Machine Learning*. 2006.

Bonnet, G. Sur certaines propriétés statistiques defonctions alétoires issues de transformations nonlinéaires. *Comptes Rendus Hebdomadaires des S éancesde l'Académie des Sciences*, pp. 4917–4920, 1964.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

Bui, T. D., Yan, J., and Turner, R. E. A unifying framework for gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research*, 18:1–72, 2017.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Ciolacu, M., Tehrani, A. F., Beer, R., and Popp, H. Education 4.0 – fostering student's performance with machine learning methods. In *2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME)*, 2017.

Duncker, L., Bohner, G., Boussard, J., and Sahani, M. Learning interpretable continuous-time models of latent stochastic dynamical systems. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97. PMLR, 2019.

Eraker, B. Mcmc analysis of diffusion models with application to finance. *Journal of Business & Economic Statistics*, 19:177–191, 2001.

Friedrich, R., Peinke, J., Sahimi, M., and Reza Rahimi Tabar, M. Approaching complexity by stochastic methods: From biological systems to turbulence. *Physics Reports*, 2011.

García, Otero, F. P. M. Nonparametric estimation of stochastic differential equations with sparse gaussian processes. *Physical Review E.*, 2017.

Girsanov, I. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. theory of probability and its applications, 1960.

Golightly, A. and Wilkinson, D. J. Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface focus*, 1:807–820, December 2011.

Griffiths, D. F. and Higham, D. *Numerical Methods for Ordinary Differential Equations: Initial Value Problems*. Springer-Verlag, 2010.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

Heaton, J. B., Polson, N. G., and Witte, J. H. Deep learning in finance, 2018.

Hensman, J., Matthews, A. G., Filippone, M., and Ghahramani, Z. Mcmc for variationally sparse gaussian processes. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015a.

Hensman, J., Matthews, A. G. d. G., and Ghahramani, Z. Scalable variational gaussian process classification. In *Proceedings of the Eighteen International Conference on Artificial Intelligence and Statistics*. PMLR, 2015b.

Khan, M. and Lin, W. Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 878–887. PMLR, 2017.

Khan, M. E. E. Decoupled variational gaussian inference. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Li, X., Wong, T.-K. L., Chen, R. T., and Duvenaud, D. Scalable gradients for stochastic differential equations. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882. PMLR, 2020.

Manfred Opper, C. A. The variational gaussian approximation revisited. In *Neural Computation, 21*, 2009.

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18:1–6, 2017.

McGoff, K., Mukherjee, S., and Pillai, N. Statistical inference for dynamical systems: A review. *Statistics Surveys*, 9, 2015.

Ng, A. and Jordan, M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

Price, R. A useful theorem for nonlinear devices having gaussian inputs. *IRE Transactions on Information Theory*, 4:69–72, 1958.

Raskutti, G. and Mukherjee, S. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

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

Rubanova, Y., Chen, R. T. Q., and Duvenaud, D. K. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Ruttor, A., Batz, P., and Opper, M. Approximate gaussian process inference for the drift function in stochastic differential equations. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Ryder, T., Golightly, A., McGough, A. S., and Prangle, D. Black-box variational inference for stochastic differential equations. In *International Conference on Machine Learning*, pp. 4420–4429, 2018.

Salimbeni, H., Eleftheriadis, S., and Hensman, J. Natural gradients in practice: Non-conjugate variational inference in gaussian process models. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 689–697. PMLR, 09–11 Apr 2018.

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

Särkkä, S. and Sottinen, T. Application of girsanov theorem to particle filtering of discretely observed continuous-time non-linear systems, 2008.

Sauer, T. *Numerical Solution of Stochastic Differential Equations in Finance*. 2012.

Solin, A. *Stochastic Differential Equation Methods for Spatio-Temporal Gaussian Process Regression*. Aalto University, School of Science, 2016.

Titsias, M. Variational learning of inducing variables in sparse gaussian processes. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 567–574. PMLR, 16–18 Apr 2009.

Uhlenbeck, G. E. and Ornstein, L. S. On the theory of the brownian motion. *Phys. Rev.*, 1930.

van Kampen, N. G. Stochastic processes in physics and chemistry. *Elsevier*, 2007.

Wang, Fleet, H. Gaussian process dynamical models. In *Advances in Neural Information Processing Systems*, volume 18. Curran Associates, Inc., 2005.

Wilkinson, W. J. Newt, 2021. URL `https://github.com/AaltoML/Newt`.

Errata

July 2022

1. Equations formatted using bold variables for better reading.
2. Integration variables in equations explicitly written for clarity.