Neural Computing Research Group
School of Engineering and Applied Science
Aston University
Birmingham B4 7ET
United Kingdom
Tel: +44 (0)121 333 4631
Fax: +44 (0)121 333 4586
http://www.ncrg.aston.ac.uk/

Derivations of Variational Gaussian Process Approximation Framework

Michail D. Vrettas
vrettasm@aston.ac.uk

Yuan Shen
Y.Shen2@aston.ac.uk

Dan Cornford
d.cornford@aston.ac.uk

Technical Report NCRG/2008/002 March 28, 2008

Abstract

Recently, within the VISDEM[a] project, a novel variational approximation framework has been developed for inference in partially observed, continuous space-time, diffusion processes. In this technical report all the derivations of the variational framework, from the initial work, are provided in detail to help the reader better understand the framework and its assumptions.

[a]VISDEM: Variational Inference in Stochastic Dynamic Environmental Models, an EPSRC funded research project (EP/C005848/1) involving Aston, UCL, Surrey and the Met Office.

1 Introduction

Motivated by numerical weather prediction models (for a review see [5, Ch.1]) the VISDEM project is working to establish a framework for tackling inference, in the form of state estimation (data assimilation) and parameter estimation, for such large and complex models in a fully probabilistic manner. Extending advances in machine learning, VISDEM develops a variational Bayesian framework for inference in continuous-time stochastic dynamical systems [4], in the presence of, typically, discrete time observations. It is anticipated that in the future this will allow more accurate estimation within such systems, possibly leading to better predictions and be computationally efficient, while making fewer restrictive approximations than other existing computationally feasible methodologies. In particular, within the VISDEM framework, emphasis will be placed on estimating unknown model parameters, as well as model state, thus making full use of the available observations. Results of the initial work, on estimating the state of a dynamical system, are presented in [1], and advances for estimating the unknown model parameters can be found in [2].

The rest of the report is organised as follows. First, in Section 2, we briefly introduce the basic setting. A general expression for the SDE that governs the diffusion process is given, as well the expressions for the posterior measure and the likelihood of the observations. Section (3), then provides the definition of the *variational free energy* as well as the approximating process, which in our case is considered to be a Gaussian process. Furthermore, all the equations from the smoothing algorithm are defined and derived accordingly. Finally, in Section (4), we provide the derivations for the parameter estimation. These parameters are: (**a**) the first two moments of the initial time state, (**b**) the drift parameter(s) and (**c**) the system noise covariance coefficient(s). The report concludes with a discussion.

2 Basic Setting

In order to fix ideas and make the derivations more clear to the reader we need first to introduce the basic setting on which the variational approximation framework is based on.

We consider a finite set of d-dimensional noisy observations $\{\mathbf{y}_m\}_{m=1}^M$, that are generated by a D-dimensional latent process $\mathbf{X}(t)$ (henceforth $\mathbf{X}_t$).

- We assume that the time evolution of this D-dimensional stochastic process $\mathbf{X}_t$, which represents the process about which we wish to make inference, is described by an Itô Stochastic Differential Equation (**SDE**):
$$d\mathbf{X}_t = \mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}_t)dt + \boldsymbol{\Sigma}^{1/2} d\mathbf{W}_t, \qquad d\mathbf{W}_t \sim \mathcal{N}(\mathbf{0}, dt\mathbf{I}) \tag{1}$$
where, $\mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}_t) \in \Re^D$ is (usually) a non-linear function, $\boldsymbol{\Sigma} = \text{diag}\{\boldsymbol{\sigma}_1^2, \boldsymbol{\sigma}_2^2, \ldots, \boldsymbol{\sigma}_D^2\}$ is the system noise covariance matrix and $\mathbf{W}_t \in \Re^D$ is the standard *Wiener process*.
- A discretized version of (1) can be provided by the Euler-Maruyama representation of a **SDE**. Hence we have:
$$\Delta\mathbf{x}_k \equiv \mathbf{x}_{k+1} - \mathbf{x}_k = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k)\Delta t + \sqrt{\Delta t \boldsymbol{\Sigma}}\boldsymbol{\epsilon}_k, \tag{2}$$
where Δt is the time increment and $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As $\Delta t \to 0$ this becomes equivalent to the continuous time version (1).
- The posterior measure, in the presence of independent and identically distributed (i.i.d.) observations is given by:
$$\frac{dp_{post}}{dp_{sde}} = \frac{1}{Z} \times \prod_{m=1}^{M} p(\mathbf{y}_m | \mathbf{X}_{t_m}), \tag{3}$$

using the Radon-Nikodym notation, where M denotes the number of noisy observations and Z is the normalisation constant, or marginal likelihood, or evidence (i.e. $Z = p(\mathbf{y}_{1:M})$).

- As usual, the multivariate Gaussian likelihood is given by:

$$p(\mathbf{y}_m|\mathbf{X}_{t_m}) = \mathcal{N}(\mathbf{y}_m|\mathbf{H}\mathbf{X}_{t_m}, \mathbf{R}), \tag{4}$$

where, $\mathbf{H} \in \Re^{d\times D}$ is a linear transformation between the latent state vector $\mathbf{X}_{t_m}$ and the observation $\mathbf{y}_m$ and $\mathbf{R} \in \Re^{d\times d}$ defines the observations noise covariance matrix.

A more thorough study and presentation of stochastic differential equations, as well as different discretisation schemes, can be found in many text-books. Here we cite three of the most commonly used [6], [7] and [3].

3 Approximate Inference

- The variational free energy, is defined as follows:

$$\mathcal{F}_{\boldsymbol{\Sigma}}(q, \boldsymbol{\theta}) = -\left\langle \ln \frac{p(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\Sigma})}{q(\mathbf{X}|\boldsymbol{\Sigma})} \right\rangle_q \tag{5}$$

where $p(\cdot)$ is the true posterior process of the system and $q(\cdot)$ is the one that we use as an approximation. Also $\mathbf{X} = \{\mathbf{X}_t, t_0 \leq t \leq t_f\}$ is the path of a continuous time D-dimensional stochastic process and $\langle\cdot\rangle_q$ indicates the expectation with respect to process $q(\cdot)$.

Alternative, we can see the variational free energy as the KL divergence between the approximate process $q(\mathbf{X})$ and the joint distribution of the latent states and the observations of the true system $p(\mathbf{Y}, \mathbf{X})$, as follows:

$$\begin{aligned}
\mathcal{F}_{\boldsymbol{\Sigma}}(q, \boldsymbol{\theta}) &= -\left\langle \ln \frac{p(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\Sigma})}{q(\mathbf{X}|\boldsymbol{\Sigma})} \right\rangle_q \\
&= -\int q(\mathbf{X}) \ln \frac{p(\mathbf{Y}, \mathbf{X})}{q(\mathbf{X})} d\mathbf{X} \\
&= \int q(\mathbf{X}) \ln \frac{q(\mathbf{X})}{p(\mathbf{Y}, \mathbf{X})} d\mathbf{X} \\
&= \mathrm{KL}[q(\mathbf{X}) \| p(\mathbf{Y}, \mathbf{X})]
\end{aligned} \tag{6}$$

where we have omitted the conditioning on the (hyper)parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ for notational simplicity.

- The free energy provides an upper bound to the negative marginal log-likelihood:

Starting with the *product rule* of probabilities we have :

$$\begin{aligned}
&p(\mathbf{X}, \mathbf{Y}) = p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y}) \Rightarrow \\
&p(\mathbf{Y}) = \frac{p(\mathbf{X}, \mathbf{Y})}{p(\mathbf{X}|\mathbf{Y})}
\end{aligned}$$

we apply the natural logarithm on both sides:

$$\begin{aligned}
&\ln p(\mathbf{Y}) = \ln \frac{p(\mathbf{X}, \mathbf{Y})}{p(\mathbf{X}|\mathbf{Y})} \Rightarrow \\
&\ln p(\mathbf{Y}) = \ln p(\mathbf{X}, \mathbf{Y}) - \ln p(\mathbf{X}|\mathbf{Y})
\end{aligned}$$

then we add and subtract the same quantity by introducing a new distribution $q(\mathbf{X})$:

$$\begin{aligned} -\ln p(\mathbf{Y}) &= \ln p(\mathbf{X}|\mathbf{Y}) - \ln p(\mathbf{X},\mathbf{Y}) \\ &= \ln p(\mathbf{X}|\mathbf{Y}) - \ln q(\mathbf{X}) - \ln p(\mathbf{X},\mathbf{Y}) + \ln q(\mathbf{X}) \\ &= \ln \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} - \ln \frac{p(\mathbf{X},\mathbf{Y})}{q(\mathbf{X})} \end{aligned}$$

multiplying both sides by $q(\mathbf{X})$ we have:

$$-q(\mathbf{X}) \ln p(\mathbf{Y}) = q(\mathbf{X}) \ln \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} - q(\mathbf{X}) \ln \frac{p(\mathbf{X},\mathbf{Y})}{q(\mathbf{X})} \Rightarrow$$

and integrating over $\mathbf{X}$ gives us:

$$\begin{aligned} -\int q(\mathbf{X}) \ln p(\mathbf{Y}) d\mathbf{X} = \int q(\mathbf{X}) \ln \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} d\mathbf{X} - \int q(\mathbf{X}) \ln \frac{p(\mathbf{X},\mathbf{Y})}{q(\mathbf{X})} d\mathbf{X} \Rightarrow \\ -\ln p(\mathbf{Y}) = \mathrm{KL}[q(\mathbf{X}) \| p(\mathbf{X},\mathbf{Y})] - \mathrm{KL}[q(\mathbf{X}) \| p(\mathbf{X}|\mathbf{Y})] \Rightarrow \end{aligned}$$

since $p(\mathbf{Y})$ has no dependency on $\mathbf{X}$ which leads to:

$$-\ln p(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \mathcal{F}_{\boldsymbol{\Sigma}}(q, \boldsymbol{\theta}) - \mathrm{KL}[q(\mathbf{X}|\boldsymbol{\Sigma}) \| p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\Sigma})] \leq \mathcal{F}_{\boldsymbol{\Sigma}}(q, \boldsymbol{\theta}) \tag{7}$$

because from the definition of the KL divergence we know that $\mathrm{KL} \geq 0$. Note we have added the conditioning on the (hyper)parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ here for later clarity.

3.1 Optimal approximate posterior process

- We define an approximate time-varying linear process, with the same diffusion coefficient as the process which we are approximating, $\boldsymbol{\Sigma}^{1/2}$:

$$d\mathbf{X}_t = \mathbf{g}(t, \mathbf{X}_t)dt + \boldsymbol{\Sigma}^{1/2} d\mathbf{W}_t, \qquad d\mathbf{W}_t \sim \mathcal{N}(\mathbf{0}, dt\mathbf{I}) \tag{8}$$

  where we assume: $\mathbf{g}(t, \mathbf{X}_t) = -\mathbf{A}(t)\mathbf{X}_t + \mathbf{b}(t)$, with $\mathbf{A}(t) \in \Re^{D \times D}$ (henceforth $\mathbf{A}_t$) and $\mathbf{b}(t) \in \Re^D$ (henceforth $\mathbf{b}_t$). Note that both parameters $\mathbf{A}_t$ and $\mathbf{b}_t$, are time dependent functions.

- The Gaussian marginal at time t is defined as follows:

$$q(\mathbf{X}_t|\boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{X}_t|\mathbf{m}(t), \mathbf{S}(t)) \tag{9}$$

  (henceforth q_t), where $\mathbf{m}(t) \in \Re^D$ (henceforth $\mathbf{m}_t$) and $\mathbf{S}(t) \in \Re^{D \times D}$ (henceforth $\mathbf{S}_t$), are respectively the marginal mean and marginal covariance at time t.

- The derivation of the free energy leads to the following result:

$$\mathcal{F}_{\boldsymbol{\Sigma}}(q, \boldsymbol{\theta}) = \int_{t_0}^{t_f} E_{sde}(t)dt + \int_{t_0}^{t_f} E_{obs}(t) \sum_n \delta(t - t_n)dt + \mathrm{KL}[q_0 \| p_0] \tag{10}$$

  where $\delta(t)$ is Dirac's delta function, $\mathrm{KL}[q_0 \| p_0]$ is a shorthand notation for $\mathrm{KL}[q(\mathbf{X}_0) \| p(\mathbf{X}_0)]$ and the energy functions are defined in equations (11) and (12) below:

Proof:
From equation (6) we have:

$$\begin{aligned}
\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) &= \mathrm{KL}[q(\mathbf{X})\|p(\mathbf{Y},\mathbf{X})] \\
&= \int q(\mathbf{X}) \ln \frac{q(\mathbf{X})}{p(\mathbf{Y},\mathbf{X})} d\mathbf{X} \\
&= \int q(\mathbf{X}) \ln \frac{q(\mathbf{X})}{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})} d\mathbf{X} \\
&= \underbrace{\int q(\mathbf{X}) \ln \frac{q(\mathbf{X})}{p(\mathbf{X})} d\mathbf{X}}_{\mathbf{(6a)}} - \underbrace{\int q(\mathbf{X}) \ln p(\mathbf{Y}|\mathbf{X}) d\mathbf{X}}_{\mathbf{(6b)}}
\end{aligned}$$

(6a) This integral is simply the KL divergence between the approximate prior process $q(\mathbf{X})$ and the true prior process $p(\mathbf{X})$ defined in (1). We can write this integral as:

$$\mathrm{KL}[q(\mathbf{X})\|p(\mathbf{X})] = \int q(\mathbf{X}) \ln \frac{q(\mathbf{X})}{p(\mathbf{X})} d\mathbf{X}$$

however to make the derivation more clear we will change the above notation to the one that follows to emphasise the discretization of the sample paths on the time interval (note a continuous time derivation is also possible).

$$\begin{aligned}
\mathrm{KL}\left[q(\mathbf{x}_{0:N})\|p(\mathbf{x}_{0:N})\right] &= \int\!\!\ldots\!\!\int q(\mathbf{x}_{0:N}) \ln \frac{q(\mathbf{x}_{0:N})}{p(\mathbf{x}_{0:N})} d\mathbf{x}_{0:N} \\
&= \int\!\!\ldots\!\!\int q(\mathbf{x}_{0:N}) \ln \frac{q(\mathbf{x}_0)\prod_{j=0}^{N-1} q(\mathbf{x}_{j+1}|\mathbf{x}_j)}{p(\mathbf{x}_0)\prod_{j=0}^{N-1} p(\mathbf{x}_{j+1}|\mathbf{x}_j)} d\mathbf{x}_{0:N} \\
&= \int\!\!\ldots\!\!\int q(\mathbf{x}_{0:N}) \ln \frac{q(\mathbf{x}_0)}{p(\mathbf{x}_0)} d\mathbf{x}_{0:N} + \\
&\quad \int\!\!\ldots\!\!\int q(\mathbf{x}_{0:N}) \ln \prod_{j=0}^{N-1} \left[\frac{q(\mathbf{x}_{j+1}|\mathbf{x}_j)}{p(\mathbf{x}_{j+1}|\mathbf{x}_j)}\right] d\mathbf{x}_{0:N} \\
&= \underbrace{\int q(\mathbf{x}_0) \ln \frac{q(\mathbf{x}_0)}{p(\mathbf{x}_0)} d\mathbf{x}_0}_{KL[q_0\|p_0]} + \\
&\quad \int\!\!\ldots\!\!\int q(\mathbf{x}_{0:N}) \ln \prod_{j=0}^{N-1} \left[\frac{q(\mathbf{x}_{j+1}|\mathbf{x}_j)}{p(\mathbf{x}_{j+1}|\mathbf{x}_j)}\right] d\mathbf{x}_{0:N} \\
&= \mathrm{KL}\left[q_0\|p_0\right] + \int\!\!\ldots\!\!\int q(\mathbf{x}_{0:N}) \ln \prod_{j=0}^{N-1} \left[\frac{q(\mathbf{x}_{j+1}|\mathbf{x}_j)}{p(\mathbf{x}_{j+1}|\mathbf{x}_j)}\right] d\mathbf{x}_{0:N} \\
&= \mathrm{KL}\left[q_0\|p_0\right] + \int\!\!\ldots\!\!\int q(\mathbf{x}_0) \prod_{i=0}^{N-1} q(\mathbf{x}_{i+1}|\mathbf{x}_i) \sum_{j=0}^{N-1} \ln \frac{q(\mathbf{x}_{j+1}|\mathbf{x}_j)}{p(\mathbf{x}_{j+1}|\mathbf{x}_j)} d\mathbf{x}_{0:N}
\end{aligned}$$

we arrive here from the fact that both processes are Markovian. Hence we can factorise the marginal distributions as a product of conditional distributions (i.e. the transition probabilities). That is:

$$q(\mathbf{x}_{0:N}) = q(\mathbf{x}_0) \prod_{i=0}^{N-1} q(\mathbf{x}_{i+1}|\mathbf{x}_i)$$

The same is true for $p(\mathbf{x}_{0:N})$.

Continuing our derivation we obtain:

$$\begin{aligned}
\mathrm{KL}[q||p] &= \mathrm{KL}\left[q_0\|p_0\right] + \sum_{j=0}^{N-1} \int \ldots \int \prod_{i=0}^{N-1} q(\mathbf{x}_{i+1}|\mathbf{x}_i) \ln \frac{q(\mathbf{x}_{j+1}|\mathbf{x}_j)}{p(\mathbf{x}_{j+1}|\mathbf{x}_j)} d\mathbf{x}_{1:N} \\
&= \mathrm{KL}\left[q_0\|p_0\right] + \sum_{j=0}^{N-1} \int \ldots \int \prod_{k=1}^{j} q(\mathbf{x}_k|\mathbf{x}_{k-1}) q(\mathbf{x}_{j+1}|\mathbf{x}_j) \ln \frac{q(\mathbf{x}_{j+1}|\mathbf{x}_j)}{p(\mathbf{x}_{j+1}|\mathbf{x}_j)} \times \\
&\quad \prod_{m=j+1}^{N-1} q(\mathbf{x}_{m+1}|\mathbf{x}_m) d\mathbf{x}_{1:N}
\end{aligned}$$

At this point we can make the following substitution:

$$\int \ldots \int \prod_{k=1}^{j} q(\mathbf{x}_k|\mathbf{x}_{k-1}) d\mathbf{x}_{1:j-1} = q(\mathbf{x}_j) \ ,$$

since this is equal to the marginal distribution $q(\mathbf{x}_j)$.
Hence we have:

$$\begin{aligned}
\mathrm{KL}[q||p] &= \mathrm{KL}[q_0||p_0] + \sum_{j=1}^{N-1} \int \ldots \int q(\mathbf{x}_j) q(\mathbf{x}_{j+1}|\mathbf{x}_j) \ln \frac{q(\mathbf{x}_{j+1}|\mathbf{x}_j)}{p(\mathbf{x}_{j+1}|\mathbf{x}_j)} \times \\
&\quad \prod_{m=j+1}^{N-1} q(\mathbf{x}_{m+1}|\mathbf{x}_m) d\mathbf{x}_{j:N}
\end{aligned}$$

A careful look on the right hand side, of the previous expression, after the $\left[\ln \frac{q(\mathbf{x}_{j+1}|\mathbf{x}_j)}{p(\mathbf{x}_{j+1}|\mathbf{x}_j)}\right]$, reveals a set of integrals that evaluate to one. That is:

$$\underbrace{\int q(\mathbf{x}_{j+2}|\mathbf{x}_{j+1}) d\mathbf{x}_{j+2}}_{=1} \underbrace{\int q(\mathbf{x}_{j+3}|\mathbf{x}_{j+2}) d\mathbf{x}_{j+3}}_{=1} \cdots \underbrace{\int q(\mathbf{x}_N|\mathbf{x}_{N-1}) d\mathbf{x}_N}_{=1}$$

So we are left with the following expression:

$$\begin{aligned}
\mathrm{KL}[q||p] &= \mathrm{KL}[q_0||p_0] + \sum_{j=1}^{N-1} \int q(\mathbf{x}_j) \underbrace{\int q(\mathbf{x}_{j+1}|\mathbf{x}_j) \ln \frac{q(\mathbf{x}_{j+1}|\mathbf{x}_j)}{p(\mathbf{x}_{j+1}|\mathbf{x}_j)} d\mathbf{x}_{j+1}}_{KL[q(\mathbf{x}_{j+1}|\mathbf{x}_j)||p_{sde}(\mathbf{x}_{j+1}|\mathbf{x}_j)]} d\mathbf{x}_j \\
&= \mathrm{KL}[q_0||p_0] + \sum_{j=1}^{N-1} \int q(\mathbf{x}_j) \mathbf{KL}[q(\mathbf{x}_{j+1}|\mathbf{x}_j)||p_{sde}(\mathbf{x}_{j+1}|\mathbf{x}_j)] d\mathbf{x}_j \\
&= \mathrm{KL}[q_0||p_0] + \sum_{j=1}^{N-1} \left\langle \mathbf{KL}[q(\mathbf{x}_{j+1}|\mathbf{x}_j)||p_{sde}(\mathbf{x}_{j+1}|\mathbf{x}_j)] \right\rangle_{q(x_j)}
\end{aligned}$$

The above **KL** divergence, provided that both processes p and q are Gaussians, is given by the following formula (see [9, Mathematical Appendix]):

$$\begin{aligned}
\mathrm{KL}\left[q(\mathbf{x}_{j+1}|\mathbf{x}_j)||p(\mathbf{x}_{j+1}|\mathbf{x}_j)\right] &= \frac{1}{2} \ln |\boldsymbol{\Sigma}_p \boldsymbol{\Sigma}_q^{-1}| + \\
&\quad \frac{1}{2} \mathrm{tr}\left[\boldsymbol{\Sigma}_p^{-1}\left((\mathbf{m}_p - \mathbf{m}_q)(\mathbf{m}_p - \mathbf{m}_q)^\top + \boldsymbol{\Sigma}_p - \boldsymbol{\Sigma}_q\right)\right]
\end{aligned}$$

From equations (1) and (8) we can see another critical assumption; that both processes have the same system noise covariance $\boldsymbol{\Sigma}$. Hence we can make the following substitution

the the previous expression: $\mathbf{\Sigma}_p = \mathbf{\Sigma}_q = \mathbf{\Sigma}$.

$$\begin{aligned}
\mathrm{KL}\left[q(\mathbf{x}_{j+1}|\mathbf{x}_j)||p(\mathbf{x}_{j+1}|\mathbf{x}_j)\right] &= \frac{1}{2}\ln|\mathbf{\Sigma}\mathbf{\Sigma}^{-1}| + \\
&\quad \frac{1}{2}\mathrm{tr}\left[\mathbf{\Sigma}^{-1}\Big((\mathbf{m}_p-\mathbf{m}_q)(\mathbf{m}_p-\mathbf{m}_q)^\top + \mathbf{\Sigma} - \mathbf{\Sigma}\Big)\right] \\
&= \underbrace{\frac{1}{2}\ln|\mathbf{I}|}_{=0} + \frac{1}{2}\mathrm{tr}\left[\mathbf{\Sigma}^{-1}\Big((\mathbf{m}_p-\mathbf{m}_q)(\mathbf{m}_p-\mathbf{m}_q)^\top\Big)\right] \\
&= \frac{1}{2}\mathrm{tr}\left[\mathbf{\Sigma}^{-1}\Big((\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_{j+1})-\mathbf{g}(\mathbf{x}_{j+1}))(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_{j+1})-\mathbf{g}(\mathbf{x}_{j+1})^\top)\Big)\right]\Delta t \\
&= \frac{1}{2}\Big((\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_{j+1})-\mathbf{g}(\mathbf{x}_{j+1}))^\top\mathbf{\Sigma}^{-1}(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_{j+1})-\mathbf{g}(\mathbf{x}_{j+1}))\Big)\Delta t
\end{aligned}$$

Hence for the whole discretized path p_{sde} we have:

$$\begin{aligned}
\mathrm{KL}[q||p] &= \mathrm{KL}\left[q_0\|p_0\right] + \\
&\quad \frac{1}{2}\sum_{k=1}^{N-1}\left\langle(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k)-\mathbf{g}(\mathbf{x}_k))^\top\mathbf{\Sigma}^{-1}(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k)-\mathbf{g}(\mathbf{x}_k))\right\rangle_{q_k}\Delta t
\end{aligned}$$

And in the limit of $\Delta t \to 0$ we have:

$$\mathrm{KL}[q||p_{sde}] = \frac{1}{2}\int_{t_0}^{t_f}\left\langle(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)-\mathbf{g}(t,\mathbf{X}_t))^\top\mathbf{\Sigma}^{-1}(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)-\mathbf{g}(t,\mathbf{X}_t))\right\rangle_{q_t} dt + \mathrm{KL}\left[q_0\|p_0\right]$$

The energy from the SDE is thus given by the following expression:

$$E_{sde}(t) = \frac{1}{2}\left\langle(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)-\mathbf{g}(t,\mathbf{X}_t))^\top\mathbf{\Sigma}^{-1}(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)-\mathbf{g}(t,\mathbf{X}_t))\right\rangle_{q_t} \tag{11}$$

(6b) Then we calculate the log-likelihood, noting that this is now formulated in continuous time. Hence we have:

$$\begin{aligned}
\ln p(\mathbf{Y}_t|\mathbf{X}_t) &= \ln\left(\mathcal{N}(\mathbf{Y}_t|\mathbf{H}\mathbf{X}_t,\mathbf{R})\right) \\
&= \ln\left((2\pi)^{-\frac{d}{2}}|\mathbf{R}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{Y}_t-\mathbf{H}\mathbf{X}_t)^\top\mathbf{R}^{-1}(\mathbf{Y}_t-\mathbf{H}\mathbf{X}_t)\right\}\right) \\
&= -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{R}| - \frac{1}{2}(\mathbf{Y}_t-\mathbf{H}\mathbf{X}_t)^\top\mathbf{R}^{-1}(\mathbf{Y}_t-\mathbf{H}\mathbf{X}_t)
\end{aligned}$$

Finally, we calculate the integral:

$$\begin{aligned}
\int q(\mathbf{X}_t)\ln p(\mathbf{Y}_t|\mathbf{X}_t)d\mathbf{X}_t &= -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{R}| - \\
&\quad \frac{1}{2}\int q(\mathbf{X}_t)\left((\mathbf{Y}_t-\mathbf{H}\mathbf{X}_t)^\top\mathbf{R}^{-1}(\mathbf{Y}_t-\mathbf{H}\mathbf{X}_t)\right)d\mathbf{X}_t \\
&= -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{R}| - \frac{1}{2}\left\langle(\mathbf{Y}_t-\mathbf{H}\mathbf{X}_t)^\top\mathbf{R}^{-1}(\mathbf{Y}_t-\mathbf{H}\mathbf{X}_t)\right\rangle_{q_t}
\end{aligned}$$

- The energy from the observations, at time 't', is the following:

$$E_{obs}(t) = \frac{1}{2}\left\langle(\mathbf{Y}_t-\mathbf{H}\mathbf{X}_t)^\top\mathbf{R}^{-1}(\mathbf{Y}_t-\mathbf{H}\mathbf{X}_t)\right\rangle_{q_t} + \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{R}| \tag{12}$$

where $\mathbf{Y} = \{\mathbf{Y}_t\ , t_0 \le t \le t_f\} \in \Re^d$ is a continuous-time observable process. The discrete time nature of the actual observations adds the delta function in equation (10).

3.2 Smoothing algorithm

The time evolution of the Gaussian measure (9) can be described by a set of Ordinary Differential Equations (*ODEs*). These are given in (13) and (14) and derived accordingly.

- ODEs of the means (with respect to time t):

$$\dot{\mathbf{m}}_t = -\mathbf{A}_t \mathbf{m}_t + \mathbf{b}_t \tag{13}$$

where $\dot{\mathbf{m}}_t$ is a shorthand notation for $\frac{d\mathbf{m}_t}{dt}$.
Proof:

$$\begin{aligned}
d\mathbf{m}_t &= \langle \mathbf{X}_t + d\mathbf{X}_t \rangle - \langle \mathbf{X}_t \rangle \\
&= \langle \mathbf{X}_t \rangle + \langle d\mathbf{X}_t \rangle - \langle \mathbf{X}_t \rangle \\
&= \langle d\mathbf{X}_t \rangle \\
&= \left\langle \mathbf{g}(t, \mathbf{X}_t) dt + \mathbf{\Sigma}^{1/2} d\mathbf{W}_t \right\rangle \qquad \text{(from (8))} \\
&= \langle \mathbf{g}(t, \mathbf{X}_t) \rangle \, dt + \mathbf{\Sigma}^{1/2} \underbrace{\langle d\mathbf{W}_t \rangle}_{=0} \\
&= \langle -\mathbf{A}_t \mathbf{X}_t + \mathbf{b}_t \rangle \, dt \\
&= -\mathbf{A}_t \langle \mathbf{X}_t \rangle \, dt + \mathbf{b}_t dt \\
&= -\mathbf{A}_t \mathbf{m}_t dt + \mathbf{b}_t dt
\end{aligned}$$

- ODEs of the variances (with respect to time t):

$$\dot{\mathbf{S}}_t = -\mathbf{A}_t \mathbf{S}_t - \mathbf{S}_t \mathbf{A}_t^\top + \mathbf{\Sigma} \tag{14}$$

where $\dot{\mathbf{S}}_t$ is a shorthand notation for $\frac{d\mathbf{S}_t}{dt}$.

Proof:

$$
\begin{aligned}
d\mathbf{S}_t &= \left\langle (\mathbf{X}_t - \mathbf{m}_t + d\mathbf{X}_t - d\mathbf{m}_t)(\mathbf{X}_t - \mathbf{m}_t + d\mathbf{X}_t - d\mathbf{m}_t)^\top \right\rangle - \left\langle (\mathbf{X}_t - \mathbf{m}_t)(\mathbf{X}_t - \mathbf{m}_t)^\top \right\rangle \\
&= \left\langle (\mathbf{X}_t - \mathbf{m}_t + d\mathbf{X}_t - d\mathbf{m}_t)(\mathbf{X}_t^\top - \mathbf{m}_t^\top + d\mathbf{X}_t^\top - d\mathbf{m}_t^\top) \right\rangle - \mathbf{S}_t \\
&= \left\langle \mathbf{X}_t\mathbf{X}_t^\top - \mathbf{X}_t\mathbf{m}_t^\top + \mathbf{X}_t d\mathbf{X}_t^\top - \mathbf{X}_t d\mathbf{m}_t^\top \right\rangle + \\
&\quad \left\langle -\mathbf{m}_t\mathbf{X}_t^\top + \mathbf{m}_t\mathbf{m}_t^\top - \mathbf{m}_t d\mathbf{X}_t^\top + \mathbf{m}_t d\mathbf{m}_t^\top \right\rangle + \\
&\quad \left\langle d\mathbf{X}_t\mathbf{X}_t^\top - d\mathbf{X}_t\mathbf{m}_t^\top + d\mathbf{X}_t d\mathbf{X}_t^\top - d\mathbf{X}_t d\mathbf{m}_t^\top \right\rangle + \\
&\quad \left\langle -d\mathbf{m}_t\mathbf{X}_t^\top + d\mathbf{m}_t\mathbf{m}_t^\top - d\mathbf{m}_t d\mathbf{X}_t^\top + d\mathbf{m}_t d\mathbf{m}_t^\top \right\rangle - \mathbf{S}_t \\
&= \left\langle \mathbf{X}_t\mathbf{X}_t^\top \right\rangle - \left\langle \mathbf{X}_t\mathbf{m}_t^\top \right\rangle + \left\langle \mathbf{X}_t d\mathbf{X}_t^\top \right\rangle - \left\langle \mathbf{X}_t d\mathbf{m}_t^\top \right\rangle + \\
&\quad \left\langle -\mathbf{m}_t\mathbf{X}_t^\top \right\rangle + \left\langle \mathbf{m}_t\mathbf{m}_t^\top \right\rangle - \left\langle \mathbf{m}_t d\mathbf{X}_t^\top \right\rangle + \left\langle \mathbf{m}_t d\mathbf{m}_t^\top \right\rangle + \\
&\quad \left\langle d\mathbf{X}_t\mathbf{X}_t^\top \right\rangle - \left\langle d\mathbf{X}_t\mathbf{m}_t^\top \right\rangle + \left\langle d\mathbf{X}_t d\mathbf{X}_t^\top \right\rangle - \left\langle d\mathbf{X}_t d\mathbf{m}_t^\top \right\rangle + \\
&\quad \left\langle -d\mathbf{m}_t\mathbf{X}_t^\top \right\rangle + \left\langle d\mathbf{m}_t\mathbf{m}_t^\top \right\rangle - \left\langle d\mathbf{m}_t d\mathbf{X}_t^\top \right\rangle + \left\langle d\mathbf{m}_t d\mathbf{m}_t^\top \right\rangle - \mathbf{S}_t \\
&= \mathbf{m}_t\mathbf{m}_t^\top + \mathbf{S}_t - \mathbf{m}_t\mathbf{m}_t^\top - \mathbf{m}_t\mathbf{m}_t^\top \mathbf{A}_t^\top dt - \mathbf{S}_t\mathbf{A}_t^\top dt + \mathbf{m}_t\mathbf{b}_t^\top dt + \\
&\quad \mathbf{m}_t\mathbf{m}_t^\top \mathbf{A}_t^\top dt - \mathbf{m}_t\mathbf{b}_t^\top dt - \mathbf{m}_t\mathbf{m}_t^\top + \mathbf{m}_t\mathbf{m}_t^\top + \mathbf{m}_t\mathbf{m}_t^\top \mathbf{A}_t^\top dt - \\
&\quad \mathbf{m}_t\mathbf{b}_t^\top dt - \mathbf{m}_t\mathbf{m}_t^\top \mathbf{A}_t^\top dt + \mathbf{m}_t\mathbf{b}_t^\top dt - \mathbf{A}_t\mathbf{m}_t\mathbf{m}_t^\top dt - \\
&\quad \mathbf{A}_t\mathbf{S}_t dt + \mathbf{b}_t\mathbf{m}_t^\top dt + \mathbf{A}_t\mathbf{m}_t\mathbf{m}_t^\top dt - \mathbf{b}_t\mathbf{m}_t^\top dt + \boldsymbol{\Sigma} dt + \\
&\quad \mathbf{A}_t\mathbf{m}_t\mathbf{m}_t^\top dt - \mathbf{b}_t\mathbf{m}_t^\top dt - \mathbf{A}_t\mathbf{m}_t\mathbf{m}_t^\top dt + \mathbf{b}_t\mathbf{m}_t^\top dt + \mathcal{O}(dt^2) \\
&= -\mathbf{A}_t\mathbf{S}_t dt - \mathbf{S}_t\mathbf{A}_t^\top dt + \boldsymbol{\Sigma} dt + \mathcal{O}(dt^2)
\end{aligned}
$$

where we have used the facts:

$$\left\langle \mathbf{X}_t \mathbf{X}_t^\top \right\rangle = \mathbf{m}_t \mathbf{m}_t^\top + \mathbf{S}_t \tag{15}$$

$$\left\langle \mathbf{X}_t \mathbf{m}_t^\top \right\rangle = \mathbf{m}_t \mathbf{m}_t^\top \tag{16}$$

$$\begin{aligned} \left\langle \mathbf{X}_t d\mathbf{m}_t^\top \right\rangle &= \left\langle \mathbf{X}_t (-\mathbf{A}_t \mathbf{m}_t dt + \mathbf{b}_t dt)^\top \right\rangle \\ &= \left\langle \mathbf{X}_t (-\mathbf{m}_t^\top \mathbf{A}_t^\top + \mathbf{b}_t^\top) dt \right\rangle \\ &= \left\langle -\mathbf{X}_t \mathbf{m}_t^\top \mathbf{A}_t^\top \right\rangle dt + \left\langle \mathbf{X}_t \mathbf{b}_t^\top \right\rangle dt \\ &= -\mathbf{m}_t \mathbf{m}_t^\top \mathbf{A}_t^\top dt + \mathbf{m}_t \mathbf{b}_t^\top dt \end{aligned} \tag{17}$$

$$\left\langle \mathbf{m}_t \mathbf{X}_t^\top \right\rangle = \mathbf{m}_t \mathbf{m}_t^\top \tag{18}$$

$$\left\langle \mathbf{m}_t \mathbf{m}_t^\top \right\rangle = \mathbf{m}_t \mathbf{m}_t^\top \tag{19}$$

$$\begin{aligned} \left\langle \mathbf{m}_t d\mathbf{X}_t^\top \right\rangle &= \mathbf{m}_t \left\langle d\mathbf{X}_t^\top \right\rangle \\ &= \mathbf{m}_t (-\mathbf{A}_t \mathbf{m}_t dt + \mathbf{b}_t dt)^\top \\ &= \mathbf{m}_t (-\mathbf{m}_t^\top \mathbf{A}_t^\top dt + \mathbf{b}_t^\top dt) \\ &= -\mathbf{m}_t \mathbf{m}_t^\top \mathbf{A}_t^\top dt + \mathbf{m}_t \mathbf{b}_t^\top dt \end{aligned} \tag{20}$$

$$\begin{aligned} \left\langle \mathbf{m}_t d\mathbf{m}_t^\top \right\rangle &= \mathbf{m}_t \left\langle (-\mathbf{A}_t \mathbf{m}_t + \mathbf{b}_t)^\top \right\rangle dt \\ &= \mathbf{m}_t \left\langle -\mathbf{m}_t^\top \mathbf{A}_t^\top + \mathbf{b}_t^\top \right\rangle dt \\ &= -\mathbf{m}_t \mathbf{m}_t^\top \mathbf{A}_t^\top dt + \mathbf{m}_t \mathbf{b}_t^\top dt \end{aligned} \tag{21}$$

$$\begin{aligned} \left\langle d\mathbf{X}_t d\mathbf{X}_t^\top \right\rangle &= \left\langle (\mathbf{g}(t, \mathbf{X}_t) dt + \mathbf{\Sigma}^{1/2} d\mathbf{W}_t)(\mathbf{g}(t, \mathbf{X}_t) dt + \mathbf{\Sigma}^{1/2} d\mathbf{W}_t)^\top \right\rangle \\ &= \left\langle (\mathbf{g}(t, \mathbf{X}_t) dt + \mathbf{\Sigma}^{1/2} d\mathbf{W}_t)(\mathbf{g}(t, \mathbf{X}_t)^\top dt + d\mathbf{W}_t^\top \mathbf{\Sigma}^{1/2}) \right\rangle \\ &= \underbrace{\left\langle \mathbf{g}(t, \mathbf{X}_t) \mathbf{g}(t, \mathbf{X}_t)^\top \right\rangle (dt^2)}_{\mathcal{O}(dt^2)} + \underbrace{\left\langle \mathbf{g}(t, \mathbf{X}_t) dt d\mathbf{W}_t^\top \mathbf{\Sigma}^{1/2} \right\rangle}_{=0} \\ &\quad + \underbrace{\left\langle \mathbf{\Sigma}^{1/2} d\mathbf{W}_t \mathbf{g}(t, \mathbf{X}_t)^\top dt \right\rangle}_{=0} + \left\langle \mathbf{\Sigma}^{1/2} d\mathbf{W}_t d\mathbf{W}_t^\top \mathbf{\Sigma}^{1/2} \right\rangle \\ &= \mathbf{\Sigma}^{1/2} \underbrace{\left\langle d\mathbf{W}_t d\mathbf{W}_t^\top \right\rangle}_{=\mathrm{dt}\mathbf{I}} \mathbf{\Sigma}^{1/2} + \mathcal{O}(dt^2) \\ &= \mathbf{\Sigma}^{1/2} dt \mathbf{I} \mathbf{\Sigma}^{1/2} + \mathcal{O}(dt^2) \\ &= dt \mathbf{\Sigma} + \mathcal{O}(dt^2) \end{aligned} \tag{22}$$

$$\begin{aligned} \left\langle d\mathbf{X}_t d\mathbf{m}_t^\top \right\rangle &= \left\langle (\mathbf{g}(t, \mathbf{X}_t) dt + \mathbf{\Sigma}^{1/2} d\mathbf{W}_t)(-\mathbf{A}_t \mathbf{m}_t dt + \mathbf{b}_t dt)^\top \right\rangle \\ &= 0 + \mathcal{O}(dt^2) \end{aligned} \tag{23}$$

$$\begin{aligned} \left\langle d\mathbf{m}_t d\mathbf{X}_t^\top \right\rangle &= \left\langle (-\mathbf{A}_t \mathbf{m}_t dt + \mathbf{b}_t dt)(\mathbf{g}(t, \mathbf{X}_t) dt + \mathbf{\Sigma}^{1/2} d\mathbf{W}_t)^\top \right\rangle \\ &= 0 + \mathcal{O}(dt^2) \end{aligned} \tag{24}$$

$$\begin{aligned}
\left\langle d\mathbf{m}_t d\mathbf{m}_t^\top \right\rangle &= \left\langle (-\mathbf{A}_t\mathbf{m}_t dt + \mathbf{b}_t dt)(-\mathbf{A}_t\mathbf{m}_t dt + \mathbf{b}_t dt)^\top \right\rangle \\
&= \mathcal{O}(dt^2) \qquad (25)
\end{aligned}$$

$$\begin{aligned}
\left\langle d\mathbf{m}_t \mathbf{m}_t^\top \right\rangle &= \left\langle (-\mathbf{A}_t\mathbf{m}_t dt + \mathbf{b}_t dt)\mathbf{m}_t^\top \right\rangle \\
&= -\mathbf{A}_t\mathbf{m}_t\mathbf{m}_t^\top dt + \mathbf{b}_t\mathbf{m}_t^\top dt \qquad (26)
\end{aligned}$$

$$\begin{aligned}
\left\langle d\mathbf{m}_t \mathbf{X}_t^\top \right\rangle &= \left\langle (-\mathbf{A}_t\mathbf{m}_t dt + \mathbf{b}_t dt)\mathbf{X}_t^\top \right\rangle \\
&= -\mathbf{A}_t\mathbf{m}_t\mathbf{m}_t^\top dt + \mathbf{b}_t\mathbf{m}_t^\top dt \qquad (27)
\end{aligned}$$

$$\begin{aligned}
\left\langle d\mathbf{X}_t \mathbf{m}_t^\top \right\rangle &= \left\langle d\mathbf{X}_t \right\rangle \mathbf{m}_t^\top \\
&= -\mathbf{A}_t\mathbf{m}_t\mathbf{m}_t^\top dt + \mathbf{m}_t\mathbf{m}_t^\top dt \qquad (28)
\end{aligned}$$

$$\begin{aligned}
\left\langle d\mathbf{X}_t \mathbf{X}_t^\top \right\rangle &= \left\langle (\mathbf{g}(t,\mathbf{X}_t)dt + \boldsymbol{\Sigma}^{1/2} d\mathbf{W}_t)\mathbf{X}_t^\top \right\rangle \\
&= \left\langle \mathbf{g}(t,\mathbf{X}_t)\mathbf{X}_t^\top dt + \boldsymbol{\Sigma}^{1/2} d\mathbf{W}_t \mathbf{X}_t^\top \right\rangle \\
&= \left\langle \mathbf{g}(t,\mathbf{X}_t)\mathbf{X}_t^\top \right\rangle dt + \underbrace{\boldsymbol{\Sigma}^{1/2} \left\langle d\mathbf{W}_t \mathbf{X}_t^\top \right\rangle}_{=0} \\
&= \left\langle (-\mathbf{A}X_t + \mathbf{b})X_t^\top \right\rangle dt \\
&= -\mathbf{A}_t \left\langle \mathbf{X}_t\mathbf{X}_t^\top \right\rangle dt + \mathbf{b}_t \left\langle \mathbf{X}_t^\top \right\rangle dt \\
&= -\mathbf{A}_t\mathbf{m}_t\mathbf{m}_t^\top dt - \mathbf{A}_t\mathbf{S}_t dt + \mathbf{b}_t\mathbf{m}_t^\top dt \qquad (29)
\end{aligned}$$

$$\begin{aligned}
\left\langle \mathbf{X}_t d\mathbf{X}_t^\top \right\rangle &= \left\langle \mathbf{X}_t(\mathbf{g}(t,\mathbf{X}_t)dt + \boldsymbol{\Sigma}^{1/2} d\mathbf{W}_t)^\top \right\rangle \\
&= \left\langle \mathbf{X}_t \mathbf{g}(t,\mathbf{X}_t)^\top dt + \mathbf{X}_t d\mathbf{W}_t^\top \boldsymbol{\Sigma}^{1/2} \right\rangle \\
&= \left\langle \mathbf{X}_t \mathbf{g}(t,\mathbf{X}_t)^\top \right\rangle dt + \underbrace{\left\langle \mathbf{X}_t d\mathbf{W}_t^\top \right\rangle \boldsymbol{\Sigma}^{1/2}}_{=0} \\
&= \left\langle \mathbf{X}_t(-\mathbf{A}_t\mathbf{X}_t + \mathbf{b}_t)^\top \right\rangle dt \\
&= -\left\langle \mathbf{X}_t\mathbf{X}_t^\top \right\rangle \mathbf{A}_t^\top dt + \left\langle \mathbf{X}_t \right\rangle \mathbf{b}_t^\top dt \\
&= -\mathbf{m}_t\mathbf{m}_t^\top \mathbf{A}_t^\top dt - \mathbf{S}_t\mathbf{A}_t^\top dt + \mathbf{m}_t\mathbf{b}_t^\top dt \qquad (30)
\end{aligned}$$

- In order to ensure the above constraints (13) and (14), are satisfied, we formulate the following $\mathcal{L}$agrangian.

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\Sigma}} &= \mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f} \mathrm{tr}\{\boldsymbol{\Psi}_t(\dot{\mathbf{S}}_t + \mathbf{A}_t\mathbf{S}_t + \mathbf{S}_t\mathbf{A}_t^\top - \boldsymbol{\Sigma})\}dt - \\
&\quad \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top(\dot{\mathbf{m}}_t + \mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t)dt \qquad (31)
\end{aligned}$$

  where $\boldsymbol{\lambda}_t \in \Re^D$ and $\boldsymbol{\Psi}_t \in \Re^{D\times D}$ are time dependent Lagrange multipliers, with $\boldsymbol{\Psi}_t$ being symmetric.

- Taking the functional derivative of (31) w.r.t. $\mathbf{A}_t$ we have:

$$\begin{aligned}
\nabla_{\mathbf{A}_t}\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\Sigma}} &= \nabla_{\mathbf{A}_t}\Big(\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f}\mathrm{tr}\{\boldsymbol{\Psi}_t(\dot{\mathbf{S}}_t + \mathbf{A}_t\mathbf{S}_t + \mathbf{S}_t\mathbf{A}_t^\top - \boldsymbol{\Sigma})\}dt - \\
&\quad \int_{t_0}^{t_f}\boldsymbol{\lambda}_t^\top(\dot{\mathbf{m}}_t + \mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t)dt\Big) \\
&= \nabla_{\mathbf{A}_t}\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \nabla_{\mathbf{A}_t}\int_{t_0}^{t_f}\mathrm{tr}\{\boldsymbol{\Psi}_t(\dot{\mathbf{S}}_t + \mathbf{A}_t\mathbf{S}_t + \mathbf{S}_t\mathbf{A}_t^\top - \boldsymbol{\Sigma})\}dt - \\
&\quad \nabla_{\mathbf{A}_t}\int_{t_0}^{t_f}\boldsymbol{\lambda}_t^\top(\dot{\mathbf{m}}_t + \mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t)dt \\
&= \nabla_{\mathbf{A}_t}\Big(\int_{t_0}^{t_f}E_{sde}(t)dt + \int_{t_0}^{t_f}E_{obs}(t)\sum_n\delta(t-t_n)dt + \mathrm{KL}[q_0\|p_0]\Big) - \\
&\quad 2\nabla_{\mathbf{A}_t}\int_{t_0}^{t_f}\mathrm{tr}\{\boldsymbol{\Psi}_t\mathbf{A}_t\mathbf{S}_t\}dt - \nabla_{\mathbf{A}_t}\int_{t_0}^{t_f}\boldsymbol{\lambda}_t^\top\mathbf{A}_t\mathbf{m}_t dt \\
&= \nabla_{\mathbf{A}_t}E_{sde}(t) - 2\boldsymbol{\Psi}_t\mathbf{S}_t - \boldsymbol{\lambda}_t\mathbf{m}_t^\top
\end{aligned} \tag{32}$$

  where we have used the fact that $\boldsymbol{\Psi}_t$ and $\mathbf{S}_t$ are symmetric.

- Taking the functional derivative of (31) w.r.t. $\mathbf{b}_t$ we have:

$$\begin{aligned}
\nabla_{\mathbf{b}_t}\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\Sigma}} &= \nabla_{\mathbf{b}_t}\Big(\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f}\mathrm{tr}\{\boldsymbol{\Psi}_t(\dot{\mathbf{S}}_t + \mathbf{A}_t\mathbf{S}_t + \mathbf{S}_t\mathbf{A}_t^\top - \boldsymbol{\Sigma})\}dt - \\
&\quad \int_{t_0}^{t_f}\boldsymbol{\lambda}_t^\top(\dot{\mathbf{m}}_t + \mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t)dt\Big) \\
&= \nabla_{\mathbf{b}_t}\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \nabla_{\mathbf{b}_t}\int_{t_0}^{t_f}\mathrm{tr}\{\boldsymbol{\Psi}_t(\dot{\mathbf{S}}_t + \mathbf{A}_t\mathbf{S}_t + \mathbf{S}_t\mathbf{A}_t^\top - \boldsymbol{\Sigma})\}dt - \\
&\quad \nabla_{\mathbf{b}_t}\int_{t_0}^{t_f}\boldsymbol{\lambda}_t^\top(\dot{\mathbf{m}}_t + \mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t)dt \\
&= \nabla_{\mathbf{b}_t}\Big(\int_{t_0}^{t_f}E_{sde}(t)dt + \int_{t_0}^{t_f}E_{obs}(t)\sum_n\delta(t-t_n)dt + \mathrm{KL}[q_0\|p_0]\Big) + \\
&\quad \nabla_{\mathbf{b}_t}\int_{t_0}^{t_f}\boldsymbol{\lambda}_t^\top\mathbf{b}_t dt \\
&= \nabla_{\mathbf{b}_t}E_{sde}(t) + \boldsymbol{\lambda}_t
\end{aligned} \tag{33}$$

  At this point we can derive the functional derivatives of the energy E_{sde} with respect to the variational functions $\mathbf{A}_t$ and $\mathbf{b}_t$.

- To achieve this we need to differentiate (11) w.r.t. $\mathbf{b}_t$:

$$\begin{aligned}
\nabla_{\mathbf{b}_t}E_{sde}(t) &= \nabla_{\mathbf{b}_t}\left(\frac{1}{2}\left\langle(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t))^\top\boldsymbol{\Sigma}^{-1}(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t))\right\rangle_{q_t}\right) \\
&= \frac{1}{2}\left\langle\nabla_{\mathbf{b}_t}\left[(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t))^\top\boldsymbol{\Sigma}^{-1}(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t))\right]\right\rangle_{q_t} \\
&= \frac{1}{2}\left\langle\nabla_{\mathbf{b}_t}\left[\Big((\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) + \mathbf{A}_t\mathbf{X}_t) - \mathbf{b}_t\Big)^\top\boldsymbol{\Sigma}^{-1}\Big((\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) + \mathbf{A}_t\mathbf{X}_t) - \mathbf{b}_t\Big)\right]\right\rangle_{q_t} \\
&= -\frac{1}{2}2\boldsymbol{\Sigma}^{-1}\left(\langle\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t)\rangle_{q_t}\right) \\
&= -\boldsymbol{\Sigma}^{-1}\left(\langle\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\rangle_{q_t} + \mathbf{A}_t\langle\mathbf{X}_t\rangle_{q_t} - \mathbf{b}_t\right)
\end{aligned} \tag{34}$$

- Also from (33) and (34) we have:

$$\begin{aligned}\nabla_{\mathbf{b}_t} E_{sde}(t) &= -\boldsymbol{\lambda}_t \\ \nabla_{\mathbf{b}_t} E_{sde}(t) &= -\boldsymbol{\Sigma}^{-1}\left(\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\rangle_{q_t} + \mathbf{A}_t \langle \mathbf{X}_t\rangle_{q_t} - \mathbf{b}_t\right)\end{aligned}$$

from the above equations we have:

$$\begin{aligned}&\boldsymbol{\lambda}_t = \boldsymbol{\Sigma}^{-1}\left(\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\rangle_{q_t} + \mathbf{A}_t \langle \mathbf{X}_t\rangle_{q_t} - \mathbf{b}_t\right) \Rightarrow \\ &\mathbf{b}_t = \langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\rangle_{q_t} + \mathbf{A}_t\mathbf{m}_t - \boldsymbol{\Sigma}\boldsymbol{\lambda}_t \end{aligned} \tag{35}$$

where (35) is the update variational function of $\mathbf{b}_t$.

- We follow the same procedure by differentiating (11) w.r.t. $\mathbf{A}_t$:

$$\begin{aligned}\nabla_{\mathbf{A}_t} E_{sde}(t) &= \nabla_{\mathbf{A}_t}\left(\frac{1}{2}\left\langle (\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)-\mathbf{g}(t,\mathbf{X}_t))^\top \boldsymbol{\Sigma}^{-1}(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)-\mathbf{g}(t,\mathbf{X}_t))\right\rangle_{q_t}\right) \\ &= \frac{1}{2}\left\langle \nabla_{\mathbf{A}_t}\left[\left((\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)-\mathbf{b}_t)-(-\mathbf{A}_t\mathbf{X}_t)\right)^\top \boldsymbol{\Sigma}^{-1}\left((\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)-\mathbf{b}_t)-(-\mathbf{A}_t\mathbf{X}_t)\right)\right]\right\rangle_{q_t} \\ &= \frac{1}{2}2\boldsymbol{\Sigma}^{-1}\left\langle (\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)+\mathbf{A}_t\mathbf{X}_t-\mathbf{b}_t)\mathbf{X}_t^\top\right\rangle_{q_t} \\ &= \boldsymbol{\Sigma}^{-1}\left\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\mathbf{X}_t^\top + \mathbf{A}_t\mathbf{X}_t\mathbf{X}_t^\top - \mathbf{b}_t\mathbf{X}_t^\top\right\rangle_{q_t} \\ &= \boldsymbol{\Sigma}^{-1}\left(\left\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\mathbf{X}_t^\top\right\rangle_{q_t} + \mathbf{A}_t\left\langle \mathbf{X}_t\mathbf{X}_t^\top\right\rangle_{q_t} - \mathbf{b}_t\left\langle \mathbf{X}_t^\top\right\rangle_{q_t}\right) \\ &= \boldsymbol{\Sigma}^{-1}\left(\left\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\mathbf{X}_t^\top\right\rangle_{q_t} + \mathbf{A}_t(\mathbf{m}_t\mathbf{m}_t^\top+\mathbf{S}_t) - \mathbf{b}_t\mathbf{m}_t^\top\right) \\ &= \boldsymbol{\Sigma}^{-1}\left(\left\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\mathbf{X}_t^\top\right\rangle_{q_t} + \mathbf{A}_t(\mathbf{m}_t\mathbf{m}_t^\top+\mathbf{S}_t) - \mathbf{b}_t\mathbf{m}_t^\top + \left\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\mathbf{m}_t^\top\right\rangle_{q_t} - \left\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\mathbf{m}_t^\top\right\rangle_{q_t}\right) \\ &= \boldsymbol{\Sigma}^{-1}\left(\left\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\mathbf{X}_t^\top\right\rangle_{q_t} - \left\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\mathbf{m}_t^\top\right\rangle_{q_t} + \mathbf{A}_t\mathbf{S}_t\right) - \\ &\quad \boldsymbol{\Sigma}^{-1}\left(\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\rangle_{q_t} - \mathbf{A}_t\mathbf{m}_t + \mathbf{b}_t\right)\mathbf{m}_t^\top \\ &= \boldsymbol{\Sigma}^{-1}\left(\left\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)(\mathbf{X}_t-\mathbf{m}_t)^\top\right\rangle_{q_t} + \mathbf{A}_t\mathbf{S}_t\right) - \nabla_{\mathbf{b}_t}E_{sde}(t)\mathbf{m}_t^\top \\ &= \boldsymbol{\Sigma}^{-1}\left(\langle \nabla_{\mathbf{x}_t}\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\rangle_{q_t}\mathbf{S}_t + \mathbf{A}_t\mathbf{S}_t\right) - \nabla_{\mathbf{b}_t}E_{sde}(t)\mathbf{m}_t^\top \\ &= \boldsymbol{\Sigma}^{-1}\left(\langle \nabla_{\mathbf{x}_t}\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\rangle_{q_t} + \mathbf{A}_t\right)\mathbf{S}_t - \nabla_{\mathbf{b}_t}E_{sde}(t)\mathbf{m}_t^\top \end{aligned} \tag{36}$$

where we have made use of the Equation (35) and the following identity:

$$\langle \nabla_{\mathbf{x}_t}\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\rangle_{q_t} = \left\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)(\mathbf{X}_t-\mathbf{m}_t)^\top\right\rangle_{q_t}\mathbf{S}_t^{-1} \tag{37}$$

The proof of the above identity (37) is given below:

3.2.1 Proof

$$
\begin{aligned}
\langle \nabla_{\mathbf{x}_t} \mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}_t) \rangle_{q_t} &= \int_{-\infty}^{+\infty} \nabla_{\mathbf{x}_t} \mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}_t) q(\mathbf{X}_t) d\mathbf{X}_t \\
&= \int_{-\infty}^{+\infty} \Big[\nabla_{\mathbf{x}_t} \Big(\mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}_t) q(\mathbf{X}_t) \Big) - \mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}_t) \nabla_{\mathbf{x}_t} q(\mathbf{X}_t) \Big] d\mathbf{X}_t \\
&= \underbrace{\int_{-\infty}^{+\infty} \nabla_{\mathbf{x}_t} \Big[\mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}_t) q(\mathbf{X}_t) \Big] d\mathbf{X}_t}_{I_1 = 0} + \\
&\quad \int_{-\infty}^{+\infty} \mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}_t) q(\mathbf{X}_t) \mathbf{S}_t^{-1} (\mathbf{X}_t - \mathbf{m}_t) d\mathbf{X}_t \\
&= \int_{-\infty}^{+\infty} \mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}_t) (\mathbf{X}_t - \mathbf{m}_t)^\top \mathbf{S}_t^{-1} q(\mathbf{X}_t) d\mathbf{X}_t \\
&= \big\langle \mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}_t) (\mathbf{X}_t - \mathbf{m}_t)^\top \big\rangle_{q_t} \mathbf{S}_t^{-1}
\end{aligned} \tag{38}
$$

We must note however, that in order for the above integral, I_1, to be zero we must make the strong assumption that the unknown function $\mathbf{f}_{\boldsymbol{\theta}}(t, \mathbf{X}_t)$, "moves" slower then the Gaussian approximation process $q(\mathbf{X}_t)$, as $\mathbf{X}_t \to \infty$.

- Taking the functional derivative of (31) w.r.t. $\mathbf{m}_t$ we have:

$$
\begin{aligned}
\nabla_{\mathbf{m}_t} \mathcal{L}_{\boldsymbol{\theta}, \boldsymbol{\Sigma}} &= \nabla_{\mathbf{m}_t} \Big(\mathcal{F}_{\boldsymbol{\Sigma}}(q, \boldsymbol{\theta}) - \int_{t_0}^{t_f} \mathrm{tr}\{ \boldsymbol{\Psi}_t (\dot{\mathbf{S}}_t + 2\mathbf{A}_t \mathbf{S}_t - \boldsymbol{\Sigma}) \} dt - \\
&\quad \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top (\dot{\mathbf{m}}_t + \mathbf{A}_t \mathbf{m}_t - \mathbf{b}_t) dt \Big) \\
&= \nabla_{\mathbf{m}_t} \mathcal{F}_{\boldsymbol{\Sigma}}(q, \boldsymbol{\theta}) - \nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} \mathrm{tr}\{ \boldsymbol{\Psi}_t (\dot{\mathbf{S}}_t + 2\mathbf{A}_t \mathbf{S}_t - \boldsymbol{\Sigma}) \} dt - \\
&\quad \nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top (\dot{\mathbf{m}}_t + \mathbf{A}_t \mathbf{m}_t - \mathbf{b}_t) dt \\
&= \nabla_{\mathbf{m}_t} \Big(\int_{t_0}^{t_f} E_{sde}(t) dt + \int_{t_0}^{t_f} E_{obs}(t) \sum_n \delta(t - t_n) dt + \mathrm{KL}[q_0 \| p_0] \Big) - \\
&\quad \nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top \dot{\mathbf{m}}_t + \boldsymbol{\lambda}_t^\top \mathbf{A}_t \mathbf{m}_t - \boldsymbol{\lambda}_t^\top \mathbf{b}_t dt \\
&= \nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} E_{sde}(t) dt - \nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top \dot{\mathbf{m}}_t dt - \nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top \mathbf{A}_t \mathbf{m}_t dt \\
&= \nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} E_{sde}(t) dt + \nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} \dot{\boldsymbol{\lambda}}_t^\top \mathbf{m}_t dt - \nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top \mathbf{A}_t \mathbf{m}_t dt
\end{aligned} \tag{39}
$$

Setting this equal to zero and then rearranging leads to an ODE that describes the time evolution of the Lagrange multiplier $\boldsymbol{\lambda}_t$. Hence we have:

$$
\nabla_{\mathbf{m}_t} \mathcal{L}_{\boldsymbol{\theta}, \boldsymbol{\Sigma}} = 0 \Rightarrow
$$

$$
\nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} E_{sde}(t) dt + \nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} \dot{\boldsymbol{\lambda}}_t^\top \mathbf{m}_t dt - \nabla_{\mathbf{m}_t} \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top \mathbf{A}_t \mathbf{m}_t dt = 0
$$

$$
\nabla_{\mathbf{m}_t} E_{sde}(t) + \dot{\boldsymbol{\lambda}}_t - \mathbf{A}_t^\top \boldsymbol{\lambda}_t = 0
$$

$$\dot{\boldsymbol{\lambda}}_t = -\nabla_{\mathbf{m}_t} E_{sde}(t) + \mathbf{A}_t^\top \boldsymbol{\lambda}_t \tag{40}$$

where we have used the fact (from product rule for differentiation) that:

$$\frac{d}{dt}(\boldsymbol{\lambda}_t^\top \mathbf{m}_t) = \frac{d\boldsymbol{\lambda}_t^\top}{dt}\mathbf{m}_t + \boldsymbol{\lambda}_t^\top \frac{d\mathbf{m}_t}{dt}$$

and also the assumption that at the final time, t_f, there are no consistency constraints, that is:

$$\boldsymbol{\lambda}_{t_f} = \boldsymbol{\Psi}_{t_f} = 0$$

- Working the same way as above and taking the functional derivative of (31) w.r.t. $\mathbf{S}_t$ we have:

$$\begin{aligned}
\nabla_{\mathbf{S}_t}\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\Sigma}} &= \nabla_{\mathbf{S}_t}\Big(\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f} \mathrm{tr}\{\boldsymbol{\Psi}_t(\dot{\mathbf{S}}_t + 2\mathbf{A}_t\mathbf{S}_t - \boldsymbol{\Sigma})\}dt - \\
&\quad \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top(\dot{\mathbf{m}}_t + \mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t)dt\Big) \\
&= \nabla_{\mathbf{S}_t}\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \nabla_{\mathbf{S}_t}\int_{t_0}^{t_f} \mathrm{tr}\{\boldsymbol{\Psi}_t(\dot{\mathbf{S}}_t + 2\mathbf{A}_t\mathbf{S}_t - \boldsymbol{\Sigma})\}dt \\
&= \nabla_{\mathbf{S}_t}\Big(\int_{t_0}^{t_f} E_{sde}(t)dt + \int_{t_0}^{t_f} E_{obs}(t)\sum_n \delta(t-t_n)dt + \mathrm{KL}[q_0\|p_0]\Big) - \\
&\quad \nabla_{\mathbf{S}_t}\int_{t_0}^{t_f} \mathrm{tr}\{\boldsymbol{\Psi}_t(\dot{\mathbf{S}}_t + 2\mathbf{A}_t\mathbf{S}_t\}dt \\
&= \nabla_{\mathbf{S}_t}\int_{t_0}^{t_f} E_{sde}(t)dt - \nabla_{\mathbf{S}_t}\int_{t_0}^{t_f} \mathrm{tr}\{\boldsymbol{\Psi}_t\dot{\mathbf{S}}_t\}dt - 2\nabla_{\mathbf{S}_t}\int_{t_0}^{t_f} \mathrm{tr}\{\boldsymbol{\Psi}_t\mathbf{A}_t\mathbf{S}_t\}dt \\
&= \nabla_{\mathbf{S}_t}\int_{t_0}^{t_f} E_{sde}(t)dt + \nabla_{\mathbf{S}_t}\int_{t_0}^{t_f} \mathrm{tr}\{\dot{\boldsymbol{\Psi}}_t\mathbf{S}_t\}dt - 2\nabla_{\mathbf{S}_t}\int_{t_0}^{t_f} \mathrm{tr}\{\boldsymbol{\Psi}_t\mathbf{A}_t\mathbf{S}_t\}dt
\end{aligned} \tag{41}$$

Setting this equal to zero and then rearranging leads to an ODE that describes the time evolution of the Lagrange multiplier $\boldsymbol{\Psi}_t$. Hence we have:

$$\nabla_{\mathbf{S}_t}\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\Sigma}} = 0 \Rightarrow$$

$$\nabla_{\mathbf{S}_t}\int_{t_0}^{t_f} E_{sde}(t)dt + \nabla_{\mathbf{S}_t}\int_{t_0}^{t_f} \mathrm{tr}\{\dot{\boldsymbol{\Psi}}_t\mathbf{S}_t\}dt - 2\nabla_{\mathbf{S}_t}\int_{t_0}^{t_f} \mathrm{tr}\{\boldsymbol{\Psi}_t\mathbf{A}_t\mathbf{S}_t\}dt = 0$$

$$\nabla_{\mathbf{S}_t}E_{sde}(t) + \dot{\boldsymbol{\Psi}}_t - 2\boldsymbol{\Psi}_t\mathbf{A}_t = 0$$

$$\dot{\boldsymbol{\Psi}}_t = -\nabla_{\mathbf{S}_t}E_{sde}(t) + 2\boldsymbol{\Psi}_t\mathbf{A}_t \tag{42}$$

where we have used the fact (from properties of trace differentiation) that:

$$\begin{aligned}
\frac{d}{dt}\mathrm{tr}\{\boldsymbol{\Psi}_t\mathbf{S}_t\} &= \mathrm{tr}\{\frac{d}{dt}(\boldsymbol{\Psi}_t\mathbf{S}_t)\} \\
&= \mathrm{tr}\{\frac{d\boldsymbol{\Psi}_t}{dt}\mathbf{S}_t + \boldsymbol{\Psi}_t\frac{d\mathbf{S}_t}{dt}\} \\
&= \mathrm{tr}\{\frac{d\boldsymbol{\Psi}_t}{dt}\mathbf{S}_t\} + \mathrm{tr}\{\boldsymbol{\Psi}_t\frac{d\mathbf{S}_t}{dt}\}
\end{aligned}$$

and in addition, we made the same assumption about the final time as above.

Along with the set of ordinary differential equations, (40) and (42), which describe the time evolution of the Lagrange multipliers, whenever there is an observation we apply the following ***jump-conditions***.

- First we consider the $\boldsymbol{\lambda}_t$ jump-condition which is given by the following expression:

$$\boldsymbol{\lambda}_n^+ = \boldsymbol{\lambda}_n^- - \nabla_{\mathbf{m}_t} E_{obs}(t_n)$$

Then we calculate the functional derivative of $E_{obs}(t_n)$ w.r.t. $\mathbf{m}_t$, which plays the role of the jump-amplitude:

$$\begin{aligned}
\nabla_{\mathbf{m}_t} E_{obs}(t) &= \nabla_{\mathbf{m}_t}\left(\frac{1}{2}\left\langle(\mathbf{Y}_t - \mathbf{H}\mathbf{X}_t)^\top \mathbf{R}^{-1}(\mathbf{Y}_t - \mathbf{H}\mathbf{X}_t)\right\rangle_{q_t} + \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{R}|\right) \\
&= \frac{1}{2}\nabla_{\mathbf{m}_t}\left(\left\langle(\mathbf{Y}_t - \mathbf{H}\mathbf{X}_t)^\top \mathbf{R}^{-1}(\mathbf{Y}_t - \mathbf{H}\mathbf{X}_t)\right\rangle_{q_t}\right) \\
&= \frac{1}{2}\nabla_{\mathbf{m}_t}\left(\left\langle \mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{Y}_t - \mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{X}_t - \mathbf{X}_t^\top\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{Y}_t + \mathbf{X}_t^\top\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{X}_t\right\rangle_{q_t}\right) \\
&= \frac{1}{2}\nabla_{\mathbf{m}_t}\Big(\mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{Y}_t - \mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{m}_t - \mathbf{m}_t^\top\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{Y}_t + \mathrm{tr}\left\{\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{S}_t\right\} + \\
&\quad \mathbf{m}_t^\top\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{m}_t\Big) \\
&= \frac{1}{2}\nabla_{\mathbf{m}_t}\Big(\mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{Y}_t - 2\mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{m}_t + \mathrm{tr}\left\{\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{S}_t\right\} + \mathbf{m}_t^\top\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{m}_t\Big) \\
&= \frac{1}{2}\left(-2\mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{H} + 2\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{m}_t\right) \\
&= -\mathbf{H}^\top\mathbf{R}^{-1}(\mathbf{Y}_t - \mathbf{H}\mathbf{m}_t)
\end{aligned}$$

Finally we have:

$$\boldsymbol{\lambda}(t_n^+) = \boldsymbol{\lambda}(t_n^-) + \mathbf{H}^\top\mathbf{R}^{-1}(\mathbf{Y}_{t_n} - \mathbf{H}\mathbf{m}_{t_n}) \tag{43}$$

- Then we consider the $\boldsymbol{\Psi}_t$ jump-condition which is given by the following expression:

$$\boldsymbol{\Psi}_n^+ = \boldsymbol{\Psi}_n^- - \nabla_{\mathbf{S}_t} E_{obs}(t_n)$$

Again the functional derivative of $E_{obs}(t_n)$ w.r.t. $\mathbf{S}_t$, plays the role of the jump-amplitude. Hence we have:

$$\begin{aligned}
\nabla_{\mathbf{S}_t} E_{obs}(t) &= \nabla_{\mathbf{S}_t}\left(\frac{1}{2}\left\langle(\mathbf{Y}_t - \mathbf{H}\mathbf{X}_t)^\top \mathbf{R}^{-1}(\mathbf{Y}_t - \mathbf{H}\mathbf{X}_t)\right\rangle_{q_t} + \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{R}|\right) \\
&= \frac{1}{2}\nabla_{\mathbf{S}_t}\left(\left\langle(\mathbf{Y}_t - \mathbf{H}\mathbf{X}_t)^\top \mathbf{R}^{-1}(\mathbf{Y}_t - \mathbf{H}\mathbf{X}_t)\right\rangle_{q_t}\right) \\
&= \frac{1}{2}\nabla_{\mathbf{S}_t}\left(\left\langle \mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{Y}_t - \mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{X}_t - \mathbf{X}_t^\top\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{Y}_t + \mathbf{X}_t^\top\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{X}_t\right\rangle_{q_t}\right) \\
&= \frac{1}{2}\nabla_{\mathbf{S}_t}\Big(\mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{Y}_t - \mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{m}_t - \mathbf{m}_t^\top\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{Y}_t + \mathrm{tr}\left\{\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{S}_t\right\} + \\
&\quad \mathbf{m}_t^\top\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{m}_t\Big) \\
&= \frac{1}{2}\nabla_{\mathbf{S}_t}\Big(\mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{Y}_t - 2\mathbf{Y}_t^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{m}_t + \mathrm{tr}\left\{\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{S}_t\right\} + \mathbf{m}_t^\top\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{m}_t\Big) \\
&= \frac{1}{2}\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}
\end{aligned}$$

Finally we have:

$$\boldsymbol{\Psi}(t_n^+) = \boldsymbol{\Psi}(t_n^-) - \frac{1}{2}\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H} \tag{44}$$

4 Parameter Estimation

Before computing the necessary gradients for estimating the parameters we need to integrate (31) by parts in order to make the boundary conditions explicit. Hence we have:

- Integrating (31) by parts leads to the following expression:

$$
\begin{aligned}
\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\Sigma}} &= \mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f} \mathrm{tr}\Big\{\boldsymbol{\Psi}_t\Big(\dot{\mathbf{S}}_t + 2\mathbf{A}_t\mathbf{S}_t - \boldsymbol{\Sigma}\Big)\Big\}dt - \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top\Big(\dot{\mathbf{m}}_t + \mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t\Big)dt \\
&= \mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f} \mathrm{tr}\Big\{\boldsymbol{\Psi}_t\dot{\mathbf{S}}_t\Big\} + \mathrm{tr}\Big\{2\boldsymbol{\Psi}_t\mathbf{A}_t\mathbf{S}_t\Big\} - \mathrm{tr}\Big\{\boldsymbol{\Psi}_t\boldsymbol{\Sigma}\Big\}dt - \\
&\quad \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top\dot{\mathbf{m}}_t + \boldsymbol{\lambda}_t^\top\mathbf{A}_t\mathbf{m}_t - \boldsymbol{\lambda}_t^\top\mathbf{b}_t dt \\
&= \mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f} \frac{d}{dt}\mathrm{tr}\Big\{\boldsymbol{\Psi}_t\mathbf{S}_t\Big\} - \mathrm{tr}\Big\{\dot{\boldsymbol{\Psi}}_t\mathbf{S}_t\Big\} + \mathrm{tr}\Big\{2\boldsymbol{\Psi}_t\mathbf{A}_t\mathbf{S}_t\Big\} - \mathrm{tr}\Big\{\boldsymbol{\Psi}_t\boldsymbol{\Sigma}\Big\}dt - \\
&\quad \int_{t_0}^{t_f} \frac{d}{dt}(\boldsymbol{\lambda}_t^\top\mathbf{m}_t) - \dot{\boldsymbol{\lambda}}_t^\top\mathbf{m}_t + \boldsymbol{\lambda}_t^\top\mathbf{A}_t\mathbf{m}_t - \boldsymbol{\lambda}_t^\top\mathbf{b}_t dt \\
&= \mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f} \mathrm{tr}\Big\{\boldsymbol{\Psi}_t\Big(2\mathbf{A}_t\mathbf{S}_t - \boldsymbol{\Sigma}\Big) - \dot{\boldsymbol{\Psi}}_t\mathbf{S}_t\Big\}dt - \\
&\quad \int_{t_0}^{t_f} \Big\{\boldsymbol{\lambda}_t^\top\Big(\mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t\Big) - \dot{\boldsymbol{\lambda}}_t^\top\mathbf{m}_t\Big\}dt - \int_{t_0}^{t_f} \frac{d}{dt}\mathrm{tr}\Big\{\boldsymbol{\Psi}_t\mathbf{S}_t\Big\} + \frac{d}{dt}(\boldsymbol{\lambda}_t^\top\mathbf{m}_t)dt \\
&= \mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f} \mathrm{tr}\Big\{\boldsymbol{\Psi}_t\Big(2\mathbf{A}_t\mathbf{S}_t - \boldsymbol{\Sigma}\Big) - \dot{\boldsymbol{\Psi}}_t\mathbf{S}_t\Big\}dt - \\
&\quad \int_{t_0}^{t_f} \Big\{\boldsymbol{\lambda}_t^\top\Big(\mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t\Big) - \dot{\boldsymbol{\lambda}}_t^\top\mathbf{m}_t\Big\}dt - \int_{t_0}^{t_f} \frac{d}{dt}\Big(\mathrm{tr}\Big\{\boldsymbol{\Psi}_t\mathbf{S}_t\Big\} + (\boldsymbol{\lambda}_t^\top\mathbf{m}_t)\Big)dt \\
&= \mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f} \mathrm{tr}\Big\{\boldsymbol{\Psi}_t\Big(2\mathbf{A}_t\mathbf{S}_t - \boldsymbol{\Sigma}\Big) - \dot{\boldsymbol{\Psi}}_t\mathbf{S}_t\Big\}dt - \\
&\quad \int_{t_0}^{t_f} \Big\{\boldsymbol{\lambda}_t^\top\Big(\mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t\Big) - \dot{\boldsymbol{\lambda}}_t^\top\mathbf{m}_t\Big\}dt - \\
&\quad \underbrace{\boldsymbol{\lambda}_{t_f}^\top\mathbf{m}_{t_f}}_{=0} + \boldsymbol{\lambda}_{t_0}^\top\mathbf{m}_{t_0} - \underbrace{\mathrm{tr}\Big\{\boldsymbol{\Psi}_{t_f}\mathbf{S}_{t_f}\Big\}}_{=0} + \mathrm{tr}\Big\{\boldsymbol{\Psi}_{t_0}\mathbf{S}_{t_0}\Big\} \qquad (45)
\end{aligned}
$$

this arrives from the fact that at the final (algorithm) time, when the cost function has been minimised, the consistency constraints should be fulfilled. That means that both Lagrange multipliers are equal to zero.

4.1 Initial State

The initial approximate posterior process $q(\mathbf{X}_0)$ is equal to $\mathcal{N}(\mathbf{X}_0|\mathbf{m}_0,\mathbf{S}_0)$, where the initial true posterior process $p(\mathbf{X}_0)$ is chosen to be an isotropic Gaussian: $\mathcal{N}(\mathbf{X}_0|\boldsymbol{\mu}_0,\tau_0\mathbf{I})$.

- Taking the derivative of (45) with respect to $\mathbf{m}_0$ leads to the following expression:

$$\begin{aligned}
\nabla_{\mathbf{m}_0}\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\Sigma}} &= \nabla_{\mathbf{m}_0}\Bigg(\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f}\mathrm{tr}\Big\{\boldsymbol{\Psi}_t\Big(2\mathbf{A}_t\mathbf{S}_t - \boldsymbol{\Sigma}\Big) - \dot{\boldsymbol{\Psi}}_t\mathbf{S}_t\Big\}dt \\
&\quad - \int_{t_0}^{t_f}\Big\{\boldsymbol{\lambda}_t^\top\Big(\mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t\Big) - \dot{\boldsymbol{\lambda}}_t^\top\mathbf{m}_t\Big\}dt + \boldsymbol{\lambda}_0^\top\mathbf{m}_0 + \mathrm{tr}\Big\{\boldsymbol{\Psi}_0\mathbf{S}_0\Big\}\Bigg) \\
&= \nabla_{\mathbf{m}_0}\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) + \nabla_{\mathbf{m}_0}(\boldsymbol{\lambda}_0^\top\mathbf{m}_0) \\
&= \nabla_{\mathbf{m}_0}\mathrm{KL}[q(\mathbf{X}_0)||p(\mathbf{X}_0)] + \boldsymbol{\lambda}_0 \\
&= \boldsymbol{\lambda}_0 + \frac{1}{2}\nabla_{\mathbf{m}_0}\Big(\ln|\tau_0\mathbf{I}\cdot\mathbf{S}_0^{-1}| + \mathrm{tr}\Big\{(\tau_0\mathbf{I})^{-1}\left[(\mathbf{m}_0-\boldsymbol{\mu}_0)(\mathbf{m}_0-\boldsymbol{\mu}_0)^\top + \mathbf{S}_0 - \tau_0\mathbf{I}\right]\Big\}\Big) \\
&= \boldsymbol{\lambda}_0 + \frac{1}{2}\nabla_{\mathbf{m}_0}\Big(\mathrm{tr}\Big\{(\tau_0\mathbf{I})^{-1}\left[(\mathbf{m}_0-\boldsymbol{\mu}_0)(\mathbf{m}_0-\boldsymbol{\mu}_0)^\top\right]\Big\}\Big) \\
&= \boldsymbol{\lambda}_0 + \frac{1}{2}\mathrm{tr}\Big\{\nabla_{\mathbf{m}_0}\Big((\tau_0\mathbf{I})^{-1}\left[(\mathbf{m}_0-\boldsymbol{\mu}_0)(\mathbf{m}_0-\boldsymbol{\mu}_0)^\top\right]\Big)\Big\} \\
&= \boldsymbol{\lambda}_0 + \frac{1}{2}\mathrm{tr}\Big\{\nabla_{\mathbf{m}_0}\Big(\tau_0^{-1}(\mathbf{m}_0-\boldsymbol{\mu}_0)(\mathbf{m}_0-\boldsymbol{\mu}_0)^\top\Big)\Big\} \\
&= \boldsymbol{\lambda}_0 + \frac{1}{2}\mathrm{tr}\Big\{\tau_0^{-1}2(\mathbf{m}_0-\boldsymbol{\mu}_0)\Big\} \\
&= \boldsymbol{\lambda}_0 + \tau_0^{-1}(\mathbf{m}_0-\boldsymbol{\mu}_0) \qquad (46)
\end{aligned}$$

- Taking the derivative of (45) with respect to $\mathbf{S}_0$ leads to the following expression:

$$\begin{aligned}
\nabla_{\mathbf{S}_0}\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\Sigma}} &= \nabla_{\mathbf{S}_0}\Bigg(\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f}\mathrm{tr}\Big\{\boldsymbol{\Psi}_t\Big(2\mathbf{A}_t\mathbf{S}_t - \boldsymbol{\Sigma}\Big) - \dot{\boldsymbol{\Psi}}_t\mathbf{S}_t\Big\}dt \\
&\quad - \int_{t_0}^{t_f}\Big\{\boldsymbol{\lambda}_t^\top\Big(\mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t\Big) - \dot{\boldsymbol{\lambda}}_t^\top\mathbf{m}_t\Big\}dt + \boldsymbol{\lambda}_0^\top\mathbf{m}_0 + \mathrm{tr}\Big\{\boldsymbol{\Psi}_0\mathbf{S}_0\Big\}\Bigg) \\
&= \nabla_{\mathbf{S}_0}\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) + \nabla_{\mathbf{S}_0}\mathrm{tr}\Big\{\boldsymbol{\Psi}_0\mathbf{S}_0\Big\} \\
&= \nabla_{\mathbf{S}_0}\mathrm{KL}[q(\mathbf{X}_0)||p(\mathbf{X}_0)] + \boldsymbol{\Psi}_0 \\
&= \boldsymbol{\Psi}_0 + \frac{1}{2}\nabla_{\mathbf{S}_0}\Big(\ln|\tau_0\mathbf{I}\cdot\mathbf{S}_0^{-1}| + \mathrm{tr}\left\{(\tau_0\mathbf{I})^{-1}\left[(\mathbf{m}_0-\boldsymbol{\mu}_0)(\mathbf{m}_0-\boldsymbol{\mu}_0)^\top + \mathbf{S}_0 - \tau_0\mathbf{I}\right]\right\}\Big) \\
&= \boldsymbol{\Psi}_0 + \frac{1}{2}\nabla_{\mathbf{S}_0}\ln|\tau_0\mathbf{I}\cdot\mathbf{S}_0^{-1}| + \frac{1}{2}\nabla_{\mathbf{S}_0}\mathrm{tr}\left\{(\tau_0\mathbf{I})^{-1}\mathbf{S}_0\right\} \\
&= \boldsymbol{\Psi}_0 - \frac{1}{2}\mathbf{S}_0^{-1} + \frac{1}{2}(\tau_0\mathbf{I})^{-1} \\
&= \boldsymbol{\Psi}_0 + \frac{1}{2}\Big(\tau_0^{-1}\mathbf{I} - \mathbf{S}_0^{-1}\Big) \qquad (47)
\end{aligned}$$

4.2 Drift Parameter

The gradients that are associated with the drift parameters $\boldsymbol{\theta}$ depend only on the energy that comes from the SDE term in the posterior. Hence we have:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\Sigma}} &= \nabla_{\boldsymbol{\theta}}\Bigg(\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f}\mathrm{tr}\Big\{\boldsymbol{\Psi}_t\Big(2\mathbf{A}_t\mathbf{S}_t - \boldsymbol{\Sigma}\Big) - \dot{\boldsymbol{\Psi}}_t\mathbf{S}_t\Big\}dt \\
&\quad - \int_{t_0}^{t_f}\Big\{\boldsymbol{\lambda}_t^\top\Big(\mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t\Big) - \dot{\boldsymbol{\lambda}}_t^\top\mathbf{m}_t\Big\}dt + \boldsymbol{\lambda}_0^\top\mathbf{m}_0 + \mathrm{tr}\Big\{\boldsymbol{\Psi}_0\mathbf{S}_0\Big\}\Bigg) \\
&= \nabla_{\boldsymbol{\theta}}\mathcal{F}_{\boldsymbol{\Sigma}}(q,\boldsymbol{\theta}) \\
&= \nabla_{\boldsymbol{\theta}}\Bigg(\int_{t_0}^{t_f}E_{sde}(t)dt + \int_{t_0}^{t_f}E_{obs}(t)\sum_n\delta(t-t_n)dt + \mathrm{KL}[q(\mathbf{X}_0)||p(\mathbf{X}_0)]\Bigg) \\
&= \nabla_{\boldsymbol{\theta}}\int_{t_0}^{t_f}E_{sde}(t)dt \\
&= \int_{t_0}^{t_f}\nabla_{\boldsymbol{\theta}}E_{sde}(t)dt
\end{aligned} \tag{48}$$

To compute the above integral (48) we must first find the derivative of $E_{sde}(t)$ w.r.t. $\boldsymbol{\theta}$ as follows:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}E_{sde}(t) &= \nabla_{\boldsymbol{\theta}}\Bigg(\frac{1}{2}\left\langle(\mathbf{f}_{\boldsymbol{\theta}}-\mathbf{g})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{f}_{\boldsymbol{\theta}}-\mathbf{g})\right\rangle_{q_t}\Bigg) \\
&= \frac{1}{2}\left\langle\nabla_{\boldsymbol{\theta}}\left[(\mathbf{f}_{\boldsymbol{\theta}}-\mathbf{g})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{f}_{\boldsymbol{\theta}}-\mathbf{g})\right]\right\rangle_{q_t} \\
&= \frac{1}{2}\left\langle\nabla_{\boldsymbol{\theta}}\Big(\mathbf{f}_{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma}^{-1}\mathbf{f}_{\boldsymbol{\theta}} - \mathbf{f}_{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma}^{-1}\mathbf{g} - \mathbf{g}^\top\boldsymbol{\Sigma}^{-1}\mathbf{f}_{\boldsymbol{\theta}} + \mathbf{g}^\top\boldsymbol{\Sigma}^{-1}\mathbf{g}\Big)\right\rangle_{q_t} \\
&= \frac{1}{2}\left\langle\nabla_{\boldsymbol{\theta}}(\mathbf{f}_{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma}^{-1}\mathbf{f}_{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}}(\mathbf{f}_{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma}^{-1}\mathbf{g}) - \nabla_{\boldsymbol{\theta}}(\mathbf{g}^\top\boldsymbol{\Sigma}^{-1}\mathbf{f}_{\boldsymbol{\theta}})\right\rangle_{q_t} \\
&= \frac{1}{2}\left\langle(\nabla_{\boldsymbol{\theta}}\mathbf{f}_{\boldsymbol{\theta}}^\top)\boldsymbol{\Sigma}^{-1}\mathbf{f}_{\boldsymbol{\theta}} + \mathbf{f}_{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma}^{-1}(\nabla_{\boldsymbol{\theta}}\mathbf{f}_{\boldsymbol{\theta}}) - \mathbf{g}^\top\boldsymbol{\Sigma}^{-1}(\nabla_{\boldsymbol{\theta}}\mathbf{f}_{\boldsymbol{\theta}}) - \mathbf{g}^\top\boldsymbol{\Sigma}^{-1}(\nabla_{\boldsymbol{\theta}}\mathbf{f}_{\boldsymbol{\theta}})\right\rangle_{q_t} \\
&= \frac{1}{2}\left\langle 2\mathbf{f}_{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma}^{-1}(\nabla_{\boldsymbol{\theta}}\mathbf{f}_{\boldsymbol{\theta}}) - 2\mathbf{g}^\top\boldsymbol{\Sigma}^{-1}(\nabla_{\boldsymbol{\theta}}\mathbf{f}_{\boldsymbol{\theta}})\right\rangle_{q_t} \\
&= \left\langle\Big(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t)\Big)^\top\boldsymbol{\Sigma}^{-1}\Big(\nabla_{\boldsymbol{\theta}}\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\Big)\right\rangle_{q_t}
\end{aligned} \tag{49}$$

where we have used the shorthand notations $\mathbf{f}_{\boldsymbol{\theta}}$ for $\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)$ and $\mathbf{g}$ instead of $\mathbf{g}(t,\mathbf{X}_t)$.

4.3 System Noise Covariance Parameter

The estimation of the system noise is of great importance because the system noise along with the drift parameter determines the dynamics of the system.

- The gradient of (45) with respect to the system noise covariance $\mathbf{\Sigma}$ is given by:

$$\begin{aligned}
\nabla_{\mathbf{\Sigma}}\mathcal{L}_{\boldsymbol{\theta},\mathbf{\Sigma}} &= \nabla_{\mathbf{\Sigma}}\Bigg(\mathcal{F}_{\mathbf{\Sigma}}(q,\boldsymbol{\theta}) - \int_{t_0}^{t_f}\mathrm{tr}\Big\{\mathbf{\Psi}_t\Big(2\mathbf{A}_t\mathbf{S}_t - \mathbf{\Sigma}\Big) - \dot{\mathbf{\Psi}}_t\mathbf{S}_t\Big\}dt \\
&\quad - \int_{t_0}^{t_f}\Big\{\boldsymbol{\lambda}_t^\top\Big(\mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t\Big) - \dot{\boldsymbol{\lambda}}_t^\top\mathbf{m}_t\Big\}dt + \boldsymbol{\lambda}_0^\top\mathbf{m}_0 + \mathrm{tr}\Big\{\mathbf{\Psi}_0\mathbf{S}_0\Big\}\Bigg) \\
&= \nabla_{\mathbf{\Sigma}}\mathcal{F}_{\mathbf{\Sigma}}(q,\boldsymbol{\theta}) + \nabla_{\mathbf{\Sigma}}\int_{t_0}^{t_f}\mathrm{tr}\Big\{\mathbf{\Psi}_t\mathbf{\Sigma}\Big\}dt \\
&= \int_{t_0}^{t_f}\nabla_{\mathbf{\Sigma}}E_{sde}(t)dt + \int_{t_0}^{t_f}\nabla_{\mathbf{\Sigma}}\mathrm{tr}\Big\{\mathbf{\Psi}_t\mathbf{\Sigma}\Big\}dt \\
&= \int_{t_0}^{t_f}\nabla_{\mathbf{\Sigma}}E_{sde}(t)dt + \int_{t_0}^{t_f}\mathbf{\Psi}_t dt
\end{aligned}$$

and the gradient of E_{sde} with respect to $\mathbf{\Sigma}$ is given by:

$$\begin{aligned}
\nabla_{\mathbf{\Sigma}}E_{sde}(t) &= \nabla_{\mathbf{\Sigma}}\left[\frac{1}{2}\left\langle(\mathbf{f}_{\boldsymbol{\theta}} - \mathbf{g})^\top\mathbf{\Sigma}^{-1}(\mathbf{f}_{\boldsymbol{\theta}} - \mathbf{g})\right\rangle_{q_t}\right] \\
&= \frac{1}{2}\left\langle\nabla_{\mathbf{\Sigma}}\left[(\mathbf{f}_{\boldsymbol{\theta}} - \mathbf{g})^\top\mathbf{\Sigma}^{-1}(\mathbf{f}_{\boldsymbol{\theta}} - \mathbf{g})\right]\right\rangle_{q_t} \\
&= -\frac{1}{2}\left\langle\mathbf{\Sigma}^{-\top}(\mathbf{f}_{\boldsymbol{\theta}} - \mathbf{g})(\mathbf{f}_{\boldsymbol{\theta}} - \mathbf{g})^\top\mathbf{\Sigma}^{-\top}\right\rangle_{q_t} \\
&= -\frac{1}{2}\mathbf{\Sigma}^{-1}\left\langle(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t))(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t))^\top\right\rangle_{q_t}\mathbf{\Sigma}^{-1} \qquad (50)
\end{aligned}$$

because matrix $\mathbf{\Sigma}$ is symmetric.

5 Discussion and Conclusions

This technical report presents the derivation of the equations necessary for the formulation of the variational Gaussian process approximations to the posterior distribution over paths for a partially observed diffusion process. The algorithms to implement these equations will be described elsewhere, and it is anticipated that several sub-optimal variants of this method will be developed in future work, to scale the framework to realistic sized systems. In particular the work on the inference of the hyper-parameters using marginal likelihood could be extended to accommodate Bayesian treatments using a variety of frameworks including variational Bayes estimates and simpler maximum *a posteriori* estimates. The work remains in its infancy and many more empirical experiments will be required to validate the method across a range of systems.

Acknowledgements

Finally, in order for many of the derivations to become more analytically clear the use and help of Matrix Cookbook [8], is acknowledged. Also the mathematical appendix of [9], was used for the expressions of the KL divergence between two Gaussian processes.

A Special Case

A.1 Double Well System

As a special case of the derivations provided, we consider the one dimensional *double-well* system. This is a highly non-linear dynamical system that has two stable states ($\pm\,\boldsymbol{\theta}$). The drift function of this system is:

$$f_{\boldsymbol{\theta}}(X_t) = 4X_t(\theta - X_t^2), \quad \theta > 0, \tag{51}$$

The energy from the stochastic differential equation (11), is:

$$\begin{aligned}
E_{sde}(t) &= \frac{1}{2}\left\langle (\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t))^\top \boldsymbol{\Sigma}^{-1} (\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t))\right\rangle_{q_t} \\
&= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left\langle (\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t))(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t))^\top \right\rangle_{q_t} \\
&= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left\langle \left(4X_t(\theta - X_t^2) - (-A_tX_t + b_t)\right)^2 \right\rangle_{q_t} \\
&= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left\langle \left(4\theta X_t - 4X_t^3 + A_tX_t - b_t\right)^2 \right\rangle_{q_t} \\
&= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left\langle \left((4\theta + A_t)X_t - 4X_t^3 - b_t\right)^2 \right\rangle_{q_t} \\
&= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left\langle \left(c_tX_t - 4X_t^3 - b_t\right)^2 \right\rangle_{q_t}, \quad \text{by setting } c_t = (4\theta + A_t) \\
&= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left\langle c_t^2X_t^2 - 4c_tX_t^4 - b_tc_tX_t - 4c_tX_t^4 + 16X_t^6 + 4b_tX_t^3 - b_tc_tX_t + 4b_tX_t^3 + b_t^2 \right\rangle_{q_t} \\
&= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left\langle c_t^2X_t^2 - 8c_tX_t^4 - 2b_tc_tX_t + 16X_t^6 + 8b_tX_t^3 + b_t^2 \right\rangle_{q_t} \\
&= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left(c_t^2\left\langle X_t^2\right\rangle_{q_t} - 8c_t\left\langle X_t^4\right\rangle_{q_t} - 2b_tc_t\left\langle X_t\right\rangle_{q_t} + 16\left\langle X_t^6\right\rangle_{q_t} + 8b_t\left\langle X_t^3\right\rangle_{q_t} + b_t^2\right)
\end{aligned} \tag{52}$$

where $\langle X_t\rangle_{q_t}$, $\langle X_t^2\rangle_{q_t}$, $\langle X_t^3\rangle_{q_t}$, $\langle X_t^4\rangle_{q_t}$ and $\langle X_t^6\rangle_{q_t}$, are respectively the expectations of the $1st$, $2nd$, $3rd$, $4th$ and $6th$ order moments with respect to the Gaussian distribution q_t.

The gradient of E_{sde} w.r.t. the marginal means is given by:

$$\begin{aligned}
\nabla_{\mathbf{m}_t} E_{sde}(t) &= \nabla_{\mathbf{m}_t}\left(\frac{1}{2}\boldsymbol{\Sigma}^{-1}\left(c_t^2\left\langle X_t^2\right\rangle_{q_t} - 8c_t\left\langle X_t^4\right\rangle_{q_t} - 2b_tc_t\left\langle X_t\right\rangle_{q_t} + 16\left\langle X_t^6\right\rangle_{q_t} + 8b_t\left\langle X_t^3\right\rangle_{q_t} + b_t^2\right)\right) \\
&= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left(c_t^2\nabla_{\mathbf{m}_t}\left\langle X_t^2\right\rangle_{q_t} - 8c_t\nabla_{\mathbf{m}_t}\left\langle X_t^4\right\rangle_{q_t} - 2b_tc_t + 16\nabla_{\mathbf{m}_t}\left\langle X_t^6\right\rangle_{q_t} + 8b_t\nabla_{\mathbf{m}_t}\left\langle X_t^3\right\rangle_{q_t}\right)
\end{aligned} \tag{53}$$

Working in the same way, the gradient of E_{sde} w.r.t. the marginal variances is given by:

$$\begin{aligned}
\nabla_{\mathbf{S}_t} E_{sde}(t) &= \nabla_{\mathbf{S}_t}\left(\frac{1}{2}\boldsymbol{\Sigma}^{-1}\left(c_t^2\left\langle X_t^2\right\rangle_{q_t} - 8c_t\left\langle X_t^4\right\rangle_{q_t} - 2b_tc_t\left\langle X_t\right\rangle_{q_t} + 16\left\langle X_t^6\right\rangle_{q_t} + 8b_t\left\langle X_t^3\right\rangle_{q_t} + b_t^2\right)\right) \\
&= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left(c_t^2\nabla_{\mathbf{S}_t}\left\langle X_t^2\right\rangle_{q_t} - 8c_t\nabla_{\mathbf{S}_t}\left\langle X_t^4\right\rangle_{q_t} + 16\nabla_{\mathbf{S}_t}\left\langle X_t^6\right\rangle_{q_t} + 8b_t\nabla_{\mathbf{S}_t}\left\langle X_t^3\right\rangle_{q_t}\right)
\end{aligned} \tag{54}$$

The gradients of E_{sde} w.r.t. the variational parameters $\mathbf{A}_t$ and $\mathbf{b}_t$ are given by:

- From Equation (34), we have:

$$\begin{aligned}
\nabla_{\mathbf{b}_t} E_{sde}(t) &= -\boldsymbol{\Sigma}^{-1}\left(\langle \mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\rangle_{q_t} + \mathbf{A}_t \langle \mathbf{X}_t\rangle_{q_t} - \mathbf{b}_t\right) \\
&= -\boldsymbol{\Sigma}^{-1}\left(\left\langle 4X_t(\theta - X_t^2)\right\rangle_{q_t} + A_t \langle X_t\rangle_{q_t} - b_t\right) \\
&= -\boldsymbol{\Sigma}^{-1}\left(4\theta \langle X_t\rangle_{q_t} - 4\left\langle X_t^3\right\rangle_{q_t} + A_t \langle X_t\rangle_{q_t} - b_t\right) \\
&= -\boldsymbol{\Sigma}^{-1}\left((4\theta + A_t)\langle X_t\rangle_{q_t} - 4\left\langle X_t^3\right\rangle_{q_t} - b_t\right) \qquad (55)
\end{aligned}$$

- Also from Equation (36), we have:

$$\begin{aligned}
\nabla_{\mathbf{A}_t} E_{sde}(t) &= \boldsymbol{\Sigma}^{-1}\left(\langle \nabla_{\mathbf{x}_t}\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\rangle_{q_t} + \mathbf{A}_t\right)\mathbf{S}_t - \nabla_{\mathbf{b}_t} E_{sde}(t)\mathbf{m}_t^\top \\
&= \boldsymbol{\Sigma}^{-1}\left(4\theta - 12\left\langle X_t^2\right\rangle_{q_t} + A_t\right) S_t - \nabla_{\mathbf{b}_t} E_{sde}(t)\mathbf{m}_t^\top \qquad (56)
\end{aligned}$$

  where we have used the fact that:

$$\begin{aligned}
\langle \nabla_{\mathbf{x}_t}\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\rangle_{q_t} &= \left\langle \nabla_{\mathbf{x}_t}(4X_t(\theta - X_t^2))\right\rangle_{q_t} \\
&= \left\langle \nabla_{\mathbf{x}_t}(4\theta X_t - 4X_t^3)\right\rangle_{q_t} \\
&= \left\langle 4\theta - 12X_t^2)\right\rangle_{q_t} \\
&= 4\theta - 12\left\langle X_t^2\right\rangle_{q_t}
\end{aligned}$$

When estimating the (hyper)parameters we need the functional derivatives of E_{sde} with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$. These are given by:

- From Equation (49) we have:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} E_{sde}(t) &= \left\langle \left(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t)\right)^\top \boldsymbol{\Sigma}^{-1}\left(\nabla_{\boldsymbol{\theta}}\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\right)\right\rangle_{q_t} \\
&= \boldsymbol{\Sigma}^{-1}\left\langle \nabla_{\boldsymbol{\theta}}\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)\left(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t) - \mathbf{g}(t,\mathbf{X}_t)\right)^\top\right\rangle_{q_t} \\
&= \boldsymbol{\Sigma}^{-1}\left\langle 4X_t(4\theta X_t - 4X_t^3 + A_t X_t - b_t)\right\rangle_{q_t} \\
&= \boldsymbol{\Sigma}^{-1}\left\langle 16\theta X_t^2 - 16X_t^4 + 4A_t X_t^2 - 4b_t X_t\right\rangle_{q_t} \\
&= 4\boldsymbol{\Sigma}^{-1}\left\langle (4\theta + A_t)X_t^2 - 4X_t^4 - b_t X_t\right\rangle_{q_t} \\
&= 4\boldsymbol{\Sigma}^{-1}\left((4\theta + A_t)\left\langle X_t^2\right\rangle_{q_t} - 4\left\langle X_t^4\right\rangle_{q_t} - b_t\langle X_t\rangle_{q_t}\right) \qquad (57)
\end{aligned}$$

  where we have make use of the fact that:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(X_t) &= \nabla_{\boldsymbol{\theta}}\left(4X_t(\theta - X_t^2)\right) \\
&= \nabla_{\boldsymbol{\theta}}\left(4\theta X_t - 4X_t^3\right) \\
&= 4X_t
\end{aligned}$$

- Similarly, from Equation (50), we have:

$$\begin{aligned}
\nabla_{\mathbf{\Sigma}} E_{sde}(t) &= -\frac{1}{2}\mathbf{\Sigma}^{-1}\left\langle (\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)-\mathbf{g}(t,\mathbf{X}_t))(\mathbf{f}_{\boldsymbol{\theta}}(t,\mathbf{X}_t)-\mathbf{g}(t,\mathbf{X}_t))^{\top}\right\rangle_{q_t}\mathbf{\Sigma}^{-1} \\
&= -\frac{1}{2}\mathbf{\Sigma}^{-1}\left\langle \left((4\theta+A_t)X_t - 4X_t^3 - b_t\right)^2\right\rangle_{q_t}\mathbf{\Sigma}^{-1} \\
&= -\frac{1}{2}\mathbf{\Sigma}^{-1}\Big\langle c_t^2X_t^2 - 4c_tX_t^4 - b_tc_tX_t - 4c_tX_t^4 + 16X_t^6 + 4b_tX_t^3 \\
&\quad -b_tc_tX_t + 4b_tX_t^3 + b_t^2\Big\rangle_{q_t}\mathbf{\Sigma}^{-1} \\
&= -\frac{1}{2}\mathbf{\Sigma}^{-1}\left\langle c_t^2X_t^2 - 8c_tX_t^4 - 2b_tc_tX_t + 16X_t^6 + 8b_tX_t^3 + b_t^2\right\rangle_{q_t}\mathbf{\Sigma}^{-1} \\
&= -\frac{1}{2}\mathbf{\Sigma}^{-1}\Big(c_t^2\left\langle X_t^2\right\rangle_{q_t} - 8c_t\left\langle X_t^4\right\rangle_{q_t} - 2b_tc_t\left\langle X_t\right\rangle_{q_t} + 16\left\langle X_t^6\right\rangle_{q_t} \\
&\quad +8b_t\left\langle X_t^3\right\rangle_{q_t} + b_t^2\Big)\mathbf{\Sigma}^{-1} \qquad (58)
\end{aligned}$$

A.2 Gaussian Moments and related derivatives.

Uncentered moments, up to and including 8^{th} order, of a univariate Gaussian random variable X_t, were m_t and S_t are respectively the mean and variance at time 't'.

$$\begin{aligned}
\left\langle X_t^0\right\rangle_{q_t} &= 1 \\
\left\langle X_t^1\right\rangle_{q_t} &= m_t \\
\left\langle X_t^2\right\rangle_{q_t} &= m_t^2 + S_t \\
\left\langle X_t^3\right\rangle_{q_t} &= m_t^3 + 3m_tS_t \\
\left\langle X_t^4\right\rangle_{q_t} &= m_t^4 + 6m_t^2S_t + 3S_t^2 \\
\left\langle X_t^5\right\rangle_{q_t} &= m_t^5 + 10m_t^3S_t + 15m_tS_t^2 \\
\left\langle X_t^6\right\rangle_{q_t} &= m_t^6 + 15m_t^4S_t + 45m_t^2S_t^2 + 15S_t^3 \\
\left\langle X_t^7\right\rangle_{q_t} &= m_t^7 + 21m_t^5S_t + 105m_t^3S_t^2 + 105m_tS_t^3 \\
\left\langle X_t^8\right\rangle_{q_t} &= m_t^8 + 28m_t^6S + 210m_t^4S_t^2 + 420m_t^2S_t^3 + 105S_t^4
\end{aligned}$$

From here, it is easy to derive the related derivatives with respect to the means and variances at time 't'. Hence we have:

$$\begin{aligned}
\nabla_{\mathbf{m}_t} \left\langle X_t^0 \right\rangle_{q_t} &= 0 \\
\nabla_{\mathbf{m}_t} \left\langle X_t^1 \right\rangle_{q_t} &= 1 \\
\nabla_{\mathbf{m}_t} \left\langle X_t^2 \right\rangle_{q_t} &= 2m_t \\
\nabla_{\mathbf{m}_t} \left\langle X_t^3 \right\rangle_{q_t} &= 3(m_t^2 + S_t) \\
\nabla_{\mathbf{m}_t} \left\langle X_t^4 \right\rangle_{q_t} &= 4m_t^3 + 12m_t S_t \\
\nabla_{\mathbf{m}_t} \left\langle X_t^5 \right\rangle_{q_t} &= 5m_t^4 + 30m_t^2 S_t + 15S_t^2 \\
\nabla_{\mathbf{m}_t} \left\langle X_t^6 \right\rangle_{q_t} &= 6m_t^5 + 60m_t^3 S_t + 90m_t S_t^2 \\
\nabla_{\mathbf{m}_t} \left\langle X_t^7 \right\rangle_{q_t} &= 7m_t^6 + 105m_t^4 S_t + 315m_t^2 S_t^2 + 105S_t^3 \\
\nabla_{\mathbf{m}_t} \left\langle X_t^8 \right\rangle_{q_t} &= 8m_t^7 + 168m_t^5 S + 840m_t^3 S_t^2 + 840m_t S_t^3
\end{aligned}$$

Working the same way we differentiate with respect to the variances at time 't' and we obtain:

$$\begin{aligned}
\nabla_{\mathbf{S}_t} \left\langle X_t^0 \right\rangle_{q_t} &= 0 \\
\nabla_{\mathbf{S}_t} \left\langle X_t^1 \right\rangle_{q_t} &= 0 \\
\nabla_{\mathbf{S}_t} \left\langle X_t^2 \right\rangle_{q_t} &= 1 \\
\nabla_{\mathbf{S}_t} \left\langle X_t^3 \right\rangle_{q_t} &= 3m_t \\
\nabla_{\mathbf{S}_t} \left\langle X_t^4 \right\rangle_{q_t} &= 6(m_t^2 + S_t) \\
\nabla_{\mathbf{S}_t} \left\langle X_t^5 \right\rangle_{q_t} &= 10m_t^3 + 30m_t S_t \\
\nabla_{\mathbf{S}_t} \left\langle X_t^6 \right\rangle_{q_t} &= 15m_t^4 + 90m_t^2 S_t + 45S_t^2 \\
\nabla_{\mathbf{S}_t} \left\langle X_t^7 \right\rangle_{q_t} &= 21m_t^5 + 210m_t^3 S_t + 315m_t S_t^2 \\
\nabla_{\mathbf{S}_t} \left\langle X_t^8 \right\rangle_{q_t} &= 28m_t^6 + 420m_t^4 S + 1260m_t^2 S_t^2 + 420S_t^3
\end{aligned}$$

References

[1] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor. Gaussian process approximations of stochastic differential equations. *Journal of Machine Learning Research (JMLR)*, pages 1–16, 2007.

[2] C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor. Variational inference for diffusion processes. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2007.

[3] C. W. Gardiner. *Handbook of Stochastic Methods: for physics, chemistry and the natural sciences*. Springer Series in Synergetics, 3rd edition, 2003.

[4] J. Honerkamp. *Stochastic Dynamical Systems: Concepts, Numerical Methods, Data Analysis*. Wiley - VCH, 1993.

[5] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press., 2003.

[6] P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, Applications of Mathematics, 3rd edition, 1999.

[7] B. Oksendal. *Stochastic Differential Equations. An introduction with applications*. Springer-Verlag, 5th edition, 2000.

[8] K. B. Petersen and M. S. Pedersen. The Matrix Cookbook, February 2006.

[9] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 1st edition, 2006.