Chapter 1

Temporal Gaussian Processes as Stochastic Differential Equations

Consider a Gaussian process (GP) $\{f(t)\}_t$ with mean zero and covariance k. It is defined on one-dimensional $\mathbb{R}$, and the input should be intuitively understood as time (rather than space). We further assume our temporal GP is equipped with a **stationary** kernel k, i.e. we can write $k(t, t') = k(\tau)$ for $\tau := t - t'$.

Here, we will introduce and derive the stochastic differential equation (SDE) representation of such stationary temporal GPs. The rest of the chapter will go as follows: in Section 1.1, we describe the SDE needed for reformulation and present its solution; in Section 1.2, we will investigate the spectrum of the SDE solution; in Section 1.3, we will leverage the previous derivations and to find the corresponding SDE formulation of a GP with zero mean and Matern 3/2 kernel.

The material of this chapter references heavily on Solin (2016).

1.1 Stochastic Differential Equations and Their Solutions

Consider the following equation

$$a_0 f(t) + a_1 \frac{d}{dt} f(t) + a_2 \frac{d^2}{dt^2} f(t) + \cdots + a_m \frac{d^m}{dt^m} f(t) = w(t) \tag{1}$$

where $w(t)$ is a white noise process and $a_0, a_1, \ldots, a_m$ are constants. Note that by the definition of a white noise process, $w(t)$ is a Gaussian process with mean zero and covariance function

$$k_w(t, t') = \sigma^2 \delta(t - t'),$$

and it can be viewed as the derivative in time of a Brownian motion $\{B_t\}_t$.

The solution f of Equation (1) is a Gaussian process. Recall that GPs are closed under linear operations, we notice the operators applied to f on the left-hand side are all linear, whilst the right-hand side of the equation is a GP.

We will reformulate Equation (1) in matrix forms for the ease of subsequent exposition. We define

$$\boldsymbol{f}(t) = \begin{bmatrix} f(t) & \frac{d}{dt} f(t) & \cdots & \frac{d^m}{dt^m} f(t) \end{bmatrix}^T \quad \text{and} \quad \boldsymbol{w}(t) = \begin{bmatrix} w_0(t) & \cdots & w_{m-1}(t) & w(t) \end{bmatrix}^T,$$

so we have $f(t) = \boldsymbol{H}\boldsymbol{f}(t)$ and $w(t) = \boldsymbol{L} \cdot \boldsymbol{w}(t)$ for

$$\boldsymbol{H} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{L} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Subsequently, we can rewrite Equation (1) as

$$\frac{d}{dt}\boldsymbol{f}(t) = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & & & & 1 & 0 \\ -a_0 & & \cdots & \cdots & & -a_m \end{bmatrix} \boldsymbol{f}(t) + \boldsymbol{L} \cdot \boldsymbol{w}(t).$$

If we define $\boldsymbol{F}$ as the big matrix on the left to $\boldsymbol{f}(t)$ on the right-hand side, we have

$$\frac{d}{dt}\boldsymbol{f}(t) = \boldsymbol{F}\boldsymbol{f}(t) + \boldsymbol{L} \cdot \boldsymbol{w}(t). \tag{2}$$

Notice that both $\boldsymbol{F}$ and $\boldsymbol{L}$ are matrices with constant entries. We can solve the SDE (2) explicitly. Assuming we have the initial condition $\boldsymbol{f}(t')$.

First, we rewrite the equation into

$$\frac{d}{dt}\boldsymbol{f}(t) - \boldsymbol{F}\boldsymbol{f}(t) = \boldsymbol{L} \cdot \boldsymbol{w}(t)$$

and apply an integrating factor with matrix exponential to it, yielding

$$\frac{d}{dt}\exp[-\boldsymbol{F}(t-t')]\boldsymbol{f}(t) - \exp[-\boldsymbol{F}(t-t')]\boldsymbol{F}\boldsymbol{f}(t) = \exp[-\boldsymbol{F}(t-t')]\boldsymbol{L} \cdot \boldsymbol{w}(t).$$

Defining

$$g(t) := \exp[-\boldsymbol{F}(t-t')]\boldsymbol{f}(t)$$

and applying Ito lemma to it, we would obtain

$$dg(t) = \frac{d}{dt}\exp[-\boldsymbol{F}(t-t')]\boldsymbol{f}(t) - \exp[-\boldsymbol{F}(t-t')]\boldsymbol{F}\boldsymbol{f}(t),$$

which is precisely the left-hand side of the rearranged equation and that yields

$$dg(t) = \exp[-\boldsymbol{F}(t-t')]\boldsymbol{L} \cdot \boldsymbol{w}(t).$$

Integrating the above equation over t from t' to t would give us

$$\begin{aligned} g(t) - g(t') &= \int_{t'}^{t} \exp[-\boldsymbol{F}(s-t')]\boldsymbol{L}d\boldsymbol{B}_s \\ \exp[-\boldsymbol{F}(t-t')]\boldsymbol{f}(t) &= \exp[-\boldsymbol{F}(t'-t')]\boldsymbol{f}(t') + \int_{t'}^{t} \exp[-\boldsymbol{F}(s-t')]\boldsymbol{L}d\boldsymbol{B}_s \\ \boldsymbol{f}(t) &= \exp[\boldsymbol{F}(t-t')]\boldsymbol{f}(t') + \exp[\boldsymbol{F}(t-t')] \int_{t'}^{t} \exp[-\boldsymbol{F}(s-t')]\boldsymbol{L}d\boldsymbol{B}_s \\ \boldsymbol{f}(t) &= \exp[\boldsymbol{F}(t-t')]\boldsymbol{f}(t') + \int_{t'}^{t} \exp[\boldsymbol{F}(t-s)]\boldsymbol{L}d\boldsymbol{B}_s. \end{aligned}$$

By the basics of Ito integral, we can actually view $\boldsymbol{f}(t)|\boldsymbol{f}(t')$ as a Gaussian random variable

$$\boxed{\begin{aligned} &\boldsymbol{f}(t)|\boldsymbol{f}(t') \sim \boldsymbol{N}(A_t, Q_t) \\ &\qquad A_t = \exp[\boldsymbol{F}(t-t')]\boldsymbol{f}(t') \\ &\qquad Q_t = \int_{t'}^{t} \exp[\boldsymbol{F}(t-s)]\boldsymbol{L}\Sigma L^T \exp[\boldsymbol{F}^T(t-s)]ds \end{aligned}} \tag{3}$$

where Σ is the spectral density matrix of $\{\boldsymbol{w}(t)\}$ and the covariance Q_t is obtained by computing $\mathbb{E}[\boldsymbol{f}(t)\boldsymbol{f}(t)^T]$. To extract $f(t)$ would only rely on $f(t) = \boldsymbol{H}\boldsymbol{f}(t)$.

1.2 Stationary State Solution and Its Spectral Properties

The solution to Equation (2) we obtained as Equation (3) involves a Q_t covariance matrix computed via an integral. This integral induces computation costs that could be avoided in our case, which we will illustrate below.

First, we recall the white noise process $w(t)$ and consider the Fourier transform of its kernel to obtain its spectral density Σ:

$$\hat{k}(\xi) = \int_{-\infty}^{\infty} k(\tau)\exp[-2\pi i\xi\tau]d\tau = \int_{-\infty}^{\infty} \sigma^2\delta(\tau)\exp[-2\pi i\xi\tau]d\tau = \sigma^2\cdot 1\cdot 1 = \sigma^2.$$

Next, define P_t to be the covariance of $\boldsymbol{f}(t)$. We have

$$\frac{d}{dt}P_t = \frac{d}{dt}\mathbb{E}\left[\boldsymbol{f}(t)\boldsymbol{f}(t)^T\right].$$

Using the Ito lemma, we have (ignoring (t) in the notations for simplicity)

$$\begin{aligned} d\left[\boldsymbol{f}\boldsymbol{f}^T\right] &= (d\boldsymbol{f})\boldsymbol{f}^T + \boldsymbol{f}(d\boldsymbol{f})^T + (d\boldsymbol{f})(d\boldsymbol{f})^T \\ d\boldsymbol{f} &= \boldsymbol{F}\boldsymbol{f}dt + \boldsymbol{L}d\boldsymbol{B}_t \\ \mathbb{E}\left[(d\boldsymbol{f})\boldsymbol{f}^T\right] &= \boldsymbol{F}\mathbb{E}[\boldsymbol{f}\boldsymbol{f}^T]dt = \boldsymbol{F}P_t dt \\ \mathbb{E}\left[\boldsymbol{f}(d\boldsymbol{f})^T\right] &= \mathbb{E}[\boldsymbol{f}\boldsymbol{f}^T]\boldsymbol{F}dt = P_t\boldsymbol{F}^T dt \\ \mathbb{E}\left[(d\boldsymbol{f})^T(d\boldsymbol{f})^T\right] &= \boldsymbol{L}\Sigma\boldsymbol{L}^T \end{aligned}$$

using repeatedly the Ito formula in the derivations, so

$$\frac{d}{dt}P_t = \boldsymbol{F}P_t + P_t\boldsymbol{F}^T + \boldsymbol{L}\Sigma\boldsymbol{L}^T.$$

For the steady state $\boldsymbol{f}_\infty$ with its covariance P_∞, we have

$$\boldsymbol{F}P_\infty + P_\infty\boldsymbol{F}^T + \boldsymbol{L}\Sigma\boldsymbol{L}^T = 0$$

which can be used to find P_∞ given $\boldsymbol{F}, \boldsymbol{L}, \Sigma$.

Subsequently, using

$$\boldsymbol{f}(t) = \exp[\boldsymbol{F}(t-t')]\boldsymbol{f}(t') + \int_{t'}^{t}\exp[\boldsymbol{F}(t-s)]\boldsymbol{L}d\boldsymbol{B}_s,$$

we have

$$P_t = \mathbb{E}\left[\boldsymbol{f}(t)\boldsymbol{f}(t)^T\right] = \exp[\boldsymbol{F}(t-t')]P_{t'}\exp[\boldsymbol{F}^T(t-t')] + Q_t.$$

In the case where $t > t'$ and both are times beyond stationarity, we have $P_\infty = P_t = P_{t'}$ and

$$Q_t = P_\infty - A_t P_\infty A_t^T$$

using $A_t = \exp[\boldsymbol{F}(t-t')]\boldsymbol{f}(t')$.

Therefore, if we can assume the solution GP $\boldsymbol{f}$ is stationary (e.g. its kernel is stationary), we can compute P_∞ using

$$\boxed{\boldsymbol{F}P_\infty + P_\infty\boldsymbol{F}^T + \boldsymbol{L}\Sigma\boldsymbol{L}^T = 0}$$

and find

$$\boxed{Q_t = P_\infty - A_t P_\infty A_t^T}$$

for any t.

The final thing that we will explore in this section is the spectral density $S_f(\omega)$ of the $f(t)$. Revealing the spectral density of the solution $\boldsymbol{f}(t)$ (which gives us f via $\boldsymbol{H}$) will allow us to find the suitable $\boldsymbol{F}$ and Σ for the SDE representation of stationary temporal GPs with given kernels.

Applying the Fourier transform to both sides of Equation (2) gives

$$\begin{aligned}\frac{d}{dt}\boldsymbol{f}(t) &= \boldsymbol{F}\boldsymbol{f}(t) + \boldsymbol{L}\boldsymbol{w}(t)\\ \frac{d}{d\omega}\hat{\boldsymbol{f}}(\omega) &= \boldsymbol{F}\hat{\boldsymbol{f}}(\omega) + \boldsymbol{L}\hat{\boldsymbol{w}}(\omega)\\ i\omega\hat{\boldsymbol{f}}(\omega) &= \boldsymbol{F}\hat{\boldsymbol{f}}(\omega) + \boldsymbol{L}\hat{\boldsymbol{w}}(\omega)\\ \hat{\boldsymbol{f}}(\omega) &= (i\omega I - \boldsymbol{F})^{-1}\boldsymbol{L}\hat{\boldsymbol{w}}(\omega).\end{aligned}$$

The spectral density $S_f(\omega)$ of f is therefore provided by

$$\boxed{\begin{aligned}S_f(\omega) &= \mathbb{E}[\hat{f}(t)\hat{f}(t)^T]\\ &= \mathbb{E}[\boldsymbol{H}\hat{\boldsymbol{f}}(\omega)\hat{\boldsymbol{f}}(\omega)^*\boldsymbol{H}^*]\\ &= \boldsymbol{H}[(i\omega I - \boldsymbol{F})^{-1}\boldsymbol{L}]\mathbb{E}[\hat{\boldsymbol{w}}(\omega)\hat{\boldsymbol{w}}(\omega)^*][(i\omega I - \boldsymbol{F})^{-1}\boldsymbol{L}]^*\boldsymbol{H}^*\\ &= \boldsymbol{H}[(i\omega I - \boldsymbol{F})^{-1}\boldsymbol{L}]S_w(\omega)[(i\omega I - \boldsymbol{F})^{-1}\boldsymbol{L}]^*\boldsymbol{H}^*\\ &= \boldsymbol{H}[(i\omega I - \boldsymbol{F})^{-1}\boldsymbol{L}]\Sigma[(i\omega I - \boldsymbol{F})^{-1}\boldsymbol{L}]^*\boldsymbol{H}^*\end{aligned}}$$

where * is the complex conjugate operation and $S_w(\omega)$ is the spectral density of white noise w which is a constant matrix $\boldsymbol{\Sigma}$ as stated earlier.

1.3 Spectrum of Stationary GP Kernels - Matern 3/2

As alluded to earlier, by investigating the spectrum, we can choose constant matrices $\boldsymbol{F}$, Σ, P_0 such that the solution to SDE (2) is a stationary temporal GP of known kernels.

We will derive the computation for Matern 3/2 kernel

$$k(\tau) = \sigma^2\left(1 + \frac{\sqrt{3}}{l}|\tau|\right)\exp\left[-\frac{\sqrt{3}}{2}|\tau|\right]$$

which we often simplify as

$$k(\tau) = \sigma^2\left(1 + |\tau|\lambda\right)\exp\left[-\lambda|\tau|\right]$$

by defining $\lambda := \sqrt{3}/l$. The other kernels' computation can be conducted similarly, thus omitted here. One can find a list of such results in Chapter 3.3 of Solin (2016).

Firstly, the spectral density of the Matern 3/2 kernel is

$$S_k(\omega) = \frac{12\sqrt{3}\,\sigma^2/l^3}{(\lambda^2 + \omega^2)^2}.$$

The rest of the section will be devoted to the computations for figuring out suitable $\boldsymbol{F}, \Sigma, P_0$ such that $S_f = S_k$.

Consider the function

$$G(s) = \boldsymbol{H}(sI - \boldsymbol{F})^{-1}\boldsymbol{L}$$

that gives

$$S_f(\omega) = G(i\omega)\Sigma G(i\omega)^*.$$

We would wish to find a $G(s)$ of the form

$$G(s) = \frac{1}{s^2 + a_1 s + a_0},$$

that corresponds to

$$\boldsymbol{F} = \begin{bmatrix} 0 & 1\\ -a_0 & -a_1\end{bmatrix}, \qquad \boldsymbol{L} = \begin{bmatrix}0\\1\end{bmatrix}, \qquad \boldsymbol{H} = \begin{bmatrix}1 & 0\end{bmatrix}.$$

Such G is called the companion form. In the case of Matern 3/2 S_k, we would set

$$a_0 = \lambda^2, \qquad a_1 = 2\lambda.$$

Substituting the values and computing G_s using $\boldsymbol{H}(sI - \boldsymbol{F})^{-1}\boldsymbol{L}$ would verify the correctness of this choice.

Additionally, we have

$$\frac{12\sqrt{3}\,\sigma^2/l^3}{\left(\lambda^2+\omega^2\right)^2} = S_k(\omega) = S_f(\omega) = G(i\omega)\Sigma G(i\omega)^* = \frac{\Sigma}{(\lambda^2+\omega^2)^2},$$

thus $\Sigma = 12\sqrt{3}\,\sigma^2/l^3 = 4\lambda^3\sigma^2$.

It can also be solved, by appealing to

$$\boldsymbol{F}P_\infty + P_\infty\boldsymbol{F}^T + \boldsymbol{L}\Sigma\boldsymbol{L}^T = 0,$$

the steady state covariance function $P_\infty = P_0$ is

$$P_\infty = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \lambda^2\sigma^2 \end{bmatrix}.$$

Therefore, we have

$$\boxed{F = \begin{bmatrix} 0 & 1 \\ -\lambda^2 & -2\lambda \end{bmatrix}, \quad L = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 \end{bmatrix}, \qquad \Sigma = 4\lambda^3\sigma^2, \qquad P_\infty = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \lambda^2\sigma^2 \end{bmatrix}}$$

such that the solution $f(t) = \boldsymbol{H}\boldsymbol{f}(t)$ of SDE

$$\frac{d}{dt}\boldsymbol{f}(t) = \boldsymbol{F}\boldsymbol{f}(t) + \boldsymbol{L}\cdot\boldsymbol{w}(t).$$

are zero-mean GPs with Matern 3/2 kernels.

Chapter 2

Temporal Gaussian Process Regression as State Space Smoothing

After formulating a temporal GP as an SDE in Chapter 1, we will consider how one can view the task of GP regression as a smoothing of a state space model (SSM). In the case where the observations are with homoscedastic Gaussian noise, the state space model produced is a linear Gaussian model and can be filtered and smoothed using the Kalman filter and smoother.

2.1 Gaussian Process Regression

Recall from Chapter 1 that we have successfully derived the SDE formulation of temporal GPs with certain kernel choices. Gaussian process models are often used to do regression: we wish to learn an unknown function f using location-value noisy observation pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$. We would often set a Gaussian process prior $p(f)$ on f and conduct a Bayesian update of the form

$$p(f|\mathcal{D}) \propto p(f) \times p(\mathcal{D}|f)$$

where $p(\mathcal{D}|f)$ is the likelihood of observing the data given the model. Often, we would further assume the observations are obtained as

$$y_i = f(x_i) + \varepsilon_i$$

for all i where $\varepsilon_1, \dots, \varepsilon_m \sim N(0, \sigma_{\text{obs}}^2)$, making the likelihood term

$$p(\mathcal{D}|f) = \prod_{i=1}^{n} \phi\left(\frac{y_i - f(x_i)}{\sigma_{\text{obs}}}\right)$$

for standard normal density ϕ.

Under such a setup (observation with Gaussian noise), the Gaussian process regression admits conjugacy, making the update tractable.

We rewrite the observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$ for $x_i, y_i \in \mathbb{R}$ as $\mathcal{D} = \{X, y\}$ where $X, y \in \mathbb{R}^m$ and

$$y = f(X) + \varepsilon, \qquad \text{for } \varepsilon \sim N_m(0, \sigma^2 I_m).$$

Under our modelling assumptions, we could write down the (log) likelihood of the m observations y under our GP prior $f \sim \mathcal{GP}(\mu, k)$. Since $y = f(X) + \xi$, we have

$$y|X \sim N_m\left(\mu(X), k(X, X) + \sigma^2 I_m\right)$$

paramerised by θ (e.g. observation noise σ, lengthscale and variance of the kernel k) which gives us the following log likelihood

$$\log p(y|X) = -\frac{m}{2}\log(2\pi) - \log|k(X,X) + \sigma^2 I_m| - \frac{1}{2}\left(y - \mu(X)\right)^T \left(k(X,X) + \sigma^2 I_m\right)^{-1}\left(y - \mu(X)\right)$$

that we maximise w.r.t. θ to obtain the maximum likelihood estimators of the (hyper)parameters.

Next, conditional on these observations, we wish to know the distributions of the GP at test points $X_* \in \mathbb{R}^n$, i.e. the conditional distribution $y_* = f(X_*)\ |\mathcal{D}$. This can be achieved by first modelling y_* and y jointly, then conditioning on y. Using the conditional distribution formula of Section A.2, we denote for simplicity the Gram matrices

$$K = k(X,X), \qquad K_* = k(X,X_*), \qquad K_{**} = k(X_*,X_*),$$

which gives us

$$\begin{aligned} y_*\ |X_*, \mathcal{D}, \sigma^2 &\sim N_n(\mu_{y_*|\mathcal{D}}, K_{y_*|\mathcal{D}}), \\ \mu_{y_*|\mathcal{D}} &= \mu(X) + K_*^T(K + \sigma^2 I_n)^{-1}y, \\ K_{y_*|\mathcal{D}} &= K_{**} - K_*^T(K + \sigma^2 I_n)^{-1}K_*. \end{aligned}$$

In the common scenario where we assume $\mu = 0$, we further have the following **GP predictive distribution**

$$\begin{aligned} y_*\ |X_*, \mathcal{D}, \sigma^2 &\sim N_n(\mu_{y_*|\mathcal{D}}, K_{y_*|\mathcal{D}}), \\ \mu_{y_*|\mathcal{D}} &= K_*^T(K + \sigma^2 I_n)^{-1}y, \\ K_{y_*|\mathcal{D}} &= K_{**} - K_*^T(K + \sigma^2 I_n)^{-1}K_*. \end{aligned}$$

2.2 State Space Model

Consider two sets of coupled stochastic process $\{X_t\}_t$ and $\{Y_t\}_t$ where the true process of interest is driven by $\{X_t\}_t$ while we only have access to it via observation process $\{Y_t\}_t$. We assume that both processes are Markovian in the sense that

$$\begin{aligned} X_t \mid (x_{0:t-1}, y_{1:t-1}) &\sim P(\cdot|x_{t-1}) \\ Y_t \mid (x_{0:t}, y_{1:t-1}) &\sim g(\cdot|x_t) \end{aligned}$$

using the short-hand notation $x_{a:b} := (x_a, x_{a+1}, \dots, x_b)$ with $a < b$ and $a, b \in \mathbb{Z}$.

We can, graphically, portray the above process like below.

$$\begin{tikzcd}
\cdots \arrow[r, "P"] & X_{t-1} \arrow[r, "P"] \arrow[d, "g"] & X_t \arrow[r, "P"] \arrow[d, "g"] & X_{t+1} \arrow[r, "P"] \arrow[d, "g"] & \cdots & \text{(signal)} \\
\cdots & Y_{t-1} & Y_t & Y_{t+1} & \cdots & \text{(observation)}
\end{tikzcd}$$

Such a model si often called a **hidden Markov model** (HMM) or **state space model** (SSM).

There are four main tasks associated with a state space model like the one above: **predicting**, **filtering**, **smoothing**, and **parameter estimation**.

The transition kernel P and the conditional distribution g usually depend on some parameters, and we denote the full vector of parameters by θ. The parameter dependency would not be made explicit most of the time to make the notation clean. In a Bayesian framework, we can think about an SSM as a Bayesian inference problem: π_0 is the prior distribution of the signal process, and as we make further observations $y_{1:t}$, we update our belief. The likelihood functions, denoted in general by p, are

$$p(x_{0:t}) = \pi_0(x_0)\prod_{i=1}^{t} P(x_i|x_{i-1}) \qquad p(y_{1:t}|x_{0:t}) = \prod_{i=1}^{t} g(y_i|x_i).$$

We can use the Bayes formula to get the posterior distribution of the signal process $X_{0:t}$ after observing $y_{1:t}$, which is given by

$$p(x_{0:t}|y_{1:t}) = \frac{p(x_{0:t}, y_{1:t})}{p(y_{1:t})} = \frac{p(x_{0:t})p(y_{1:t}|x_{0:t})}{\int p(y_{1:t}|x_{0:t})dx_{0:t}}.$$

The distribution $p(x_{0:t}|y_{1:t})$ above is called the **smoothing distribution**, and the task of finding it is called **(complete) smoothing**. Roughly speaking, this is the task of learning the distribution of the full trajectory of the signal process given all the available data.

Sometimes, we may be only interested in knowing the distribution of the current state in the signal process instead of the whole trajectory. We wish to find the conditional distribution of X_t given observations $y_{1:t}$, which is

$$\begin{aligned} p(x_t|y_{1:t}) &= \frac{p(x_t, y_{1:t})}{p(y_{1:t})} \\ &= \frac{g(y_t|x_t)p(x_t|y_{1:t-1})p(y_{1:t-1})}{p(y_t|y_{1:t-1})p(y_{1:t-1})} \\ &= \frac{g(y_t|x_t)p(x_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} \\ &= g(y_t|x_t)\frac{\int P(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}}{\int g(y_t|x_t)p(x_t|y_{1:t-1})dx_t}. \end{aligned}$$

This distribution $p(x_t|y_{1:t})$ is called the **filtering** distribution, and the task of finding it is called **filtering**. Notice that on the right-hand side of the above equation, the integral in the denominator replies on knowing $p(x_t|y_{1:t-1})$, which is the full integral in the numerator - which depends on $p(x_{t-1}|y_{1:t-1})$. This indicates a highly iterative structure of the filtering distribution - the filtering distribution at time t depends on that at time $t-1$, which depends on that at time $t-2$, etc. Furthermore, the numerator of the right-hand side of the equation above is the **prediction distribution** $p(x_t|y_{1:t-1})$, which is essentially making the prediction of the next state in the signal distribution given all observations. The denominator $p(y_{1:t})$ of the right-hand side of the equation above is the **likelihood** of observing the data, which depends on the parameters θ. The likelihood would allow us to estimate θ, say using maximum likelihood estimation.

Therefore, we have described all four main tasks associated with an HMM. They are summarised below.

- **(predict)** Find $p(x_t|y_{1:t-1})$
- **(filter)** Find $p(x_t|y_{1:t})$
- **(smooth)** Find $p(x_{0:t}|y_{1:t})$
- **(parameter estimation)** Estimate θ using likelihood $p(y_{1:t})$

2.3 Kalman Filter and Smoothing

Consider the following system of equations

$$\begin{aligned} X_t &= \Phi X_{t-1} + \eta_t, \qquad & \eta_t &\sim N(0, B) \\ Y_t &= HX_t + \epsilon_t, \qquad & \epsilon_t &\sim N(0, R) \end{aligned} \tag{4}$$

where Φ, H, B, R are matrices that we assume to know beforehand. It can be observed that this is a special case of a state space model where P, g are chosen to be Gaussian distribution with the right mean and variance. The system described by Equation (4) is often called a **linear Gaussian model** since everything is linear and Gaussian.

Notice that the first equation of (4) is closely linked to the SDE that we discussed heavily in Chapter 1, such as that of Equations (2) and (3), by thinking X_t as GP f's value at time t. The second equation of (4) is also linked to a GP as H can be the operator that extracts f from vector $\boldsymbol{f}$ while ε is the observation noise.

In this special case, the four tasks outlined in Section 2.2 (predicting, filtering, smoothing, and parameter estimation) can be conducted in closed form due to the nice properties of Gaussians.

We define the filtering distribution $X_t|y_{1:t} \sim N(\tilde{\mu}_t, \tilde{\Sigma}_t)$ where $\tilde{\mu}_t, \tilde{\Sigma}_t$ are to be found, and the predicting distribution $X_t|y_{1:t-1} \sim N(\mu_t, \Sigma_t)$ where μ_t, Σ_t are to be found. The Gaussianity follows from the closeness of Gaussians under additions and multiplications.

From Equation (4), we know

$$X_t|x_{0:t-1}, y_{1:t-1} = \Phi x_{t-1} + \eta_t, \qquad \eta_t \sim N(0, B).$$

Similarly, using the model setup, we know

$$Y_t|y_{1:t-1}, x_{0:t} = Hx_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, R).$$

So, conditional on y_{t-1} (the other history can be omitted due to the Markovian structure of the model), we have

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} |y_{t-1} \sim N\left(\begin{bmatrix} \mu_t \\ H\mu_t \end{bmatrix}, \begin{bmatrix} \Sigma_t & \Sigma_t H^T \\ H\Sigma_t^T & H\Sigma_t H^T + R \end{bmatrix} \right)$$

using the definition that $X_t|y_{t-1} \sim N(\mu_t, \Sigma_t)$ where μ_t, Σ_t are yet to be determined. Taking the marginal of the above distribution over $Y_t = y_t$ gives, using the results of marginal multivariate Gaussians of Section A.2, the **filtering distribution**

$$\boxed{\begin{aligned} &X_t|y_{1:t} \sim N(\tilde{\mu}_t, \tilde{\Sigma}_t) \\ &\qquad \tilde{\mu}_t = \mu_t + \Sigma_t H^T (H\Sigma_t H^T + R)^{-1}(y_t - H\mu_t) =: \mu_t + K_t(y_t - H\mu_t) \\ &\qquad \tilde{\Sigma}_t = \Sigma_t - \Sigma_t H^T (H\Sigma_t H^T + R)^{-1} H\Sigma_t =: \Sigma_t - K_t H\Sigma_t \\ &\qquad K_t := \Sigma_t H^T (H\Sigma_t H^T + R)^{-1} \end{aligned}}$$

where K_t is often called the **Kalman gain**. Subsequently, we can use the propagation of X_{t+1} from X_t to derive the **prediction distribution** as

$$\boxed{\begin{aligned} &X_{t+1}|y_{1:t} \sim N(\mu_{t+1}, \Sigma_{t+1}) \\ &\qquad \mu_{t+1} = \Phi\tilde{\mu}_t \\ &\qquad \Sigma_{t+1} = \Phi\tilde{\Sigma}_t\Phi^T + B. \end{aligned}}$$

Finally, we will derive the (Rauch-Tung-Striebel, RTS) smoothing distribution $X_t|y_{1:T}$. Firstly, we have

$$\begin{bmatrix} X_t \\ X_{t+1} \end{bmatrix} |y_{1:t} \sim N\left(\begin{bmatrix} \tilde{\mu}_t \\ \mu_{t+1} \end{bmatrix}, \begin{bmatrix} \tilde{\Sigma}_t & \tilde{\Sigma}_t\Phi^T \\ \Phi\tilde{\Sigma}_t^T & \Sigma_{t+1} \end{bmatrix} \right)$$

and conditioning on X_{t+1}, using formulas in Section A.2, gives

$$\begin{aligned} X_t|X_{t+1}, y_{1:t} &\sim N\left(\tilde{\mu}_t + C_t(X_{t+1} - \mu_{t+1}), \tilde{\Sigma}_t - C_t\Sigma_{t+1}C_t^T\right) \\ C_t &:= \tilde{\Sigma}_t\Phi^T\Sigma_{t+1}^{-1}. \end{aligned}$$

Next, we consider the smoothing distribution $X_t|y_{1:T} \sim N(\hat{\mu}_t, \hat{\Sigma}_t)$ which can be obtained iteratively backwards. Notice that

$$X_T|y_{1:T} \sim N(\hat{\mu}_T, \hat{\Sigma}_T) = N(\tilde{\mu}_T, \tilde{\Sigma}_T),$$

and subsequently, we have

$$\begin{aligned} p(X_{T-1}|y_{1:T}) &= \int p(X_{T-1}|x_T, y_{1:T})p(x_T|y_{1:T})dx_T \\ X_{T-1}|y_{1:T} &\sim N\left(\tilde{\mu}_{T-1} + C_{T-1}(\hat{\mu}_T - \mu_T), \tilde{\Sigma}_{T-1} - C_{T-1}\Sigma_T C_{T-1}^T + C_{T-1}\hat{\Sigma}_t C_{T-1}^T\right) \\ C_{T-1} &:= \tilde{\Sigma}_{T-1}\Phi^T\Sigma_T^{-1}. \end{aligned}$$

Therefore, we can succinctly write

$$\begin{aligned} X_{T-1}|y_{1:T} &\sim N\left(\hat{\mu}_{T-1}, \hat{\Sigma}_{T-1}\right) \\ \hat{\mu}_{T-1} &= \tilde{\mu}_{T-1} + C_{T-1}(\hat{\mu}_T - \mu_T) \\ \hat{\Sigma}_{T-1} &= \tilde{\Sigma}_{T-1} + C_{T-1}(\hat{\Sigma}_t - \Sigma_T)C_{T-1}^T \\ C_{T-1} &:= \tilde{\Sigma}_{T-1}\Phi^T\Sigma_T^{-1}. \end{aligned}$$

and for general t, the **smoothing distribution** is given by

$$\boxed{\begin{aligned} X_t|y_{1:T} &\sim N\left(\hat{\mu}_t, \hat{\Sigma}_t\right) \\ \hat{\mu}_t &= \tilde{\mu}_t + C_t(\hat{\mu}_{t+1} - \mu_{t+1}) \\ \hat{\Sigma}_t &= \tilde{\Sigma}_t + C_t(\hat{\Sigma}_{t+1} - \Sigma_{t+1})C_t^T \\ C_t &:= \tilde{\Sigma}_t\Phi^T\Sigma_{t+1}^{-1}. \end{aligned}}$$

and C_t is often called the **smoother gain**.

One can extend the above formulas easily to accommodate for unseen time test points. When the test point t^* is after all the observed data, we will simply propagate accordingly from the last observation's filtered/smoothed distribution. If the test point is between two consecutive observations, i.e. $t_k \leq t^* \leq t_{k+1}$, we will filter it by propagating from μ_{t_k} and add the adjusted Kalman gain from y_{t_k}. A similar setup can be used to compute the smoothing distribution.

In the process of Kalman filtering, we have the observation distribution

$$Y_t|y_{1:t-1} \sim N(H\mu_t, H\Sigma_t H^T + R)$$

which has the log-likelihood

$$l(y_t) = \text{const.} - \frac{1}{2}\log\det[H\Sigma_t H^T + R] - \frac{1}{2}(y_t - H\mu_t)^T[H\Sigma_t H^T + R]^{-1}(y_t - H\mu_t).$$

Thus, let θ denote all the unknown parameters of the model, the log-likelihood of our observations $y_{1:T}$ could be computed as

$$\boxed{l(\theta; y_{1:T}) = \text{const.} + \sum_{t=1}^{T}\left[-\frac{1}{2}\log\det[H\Sigma_t H^T + R] - \frac{1}{2}(y_t - H\mu_t)^T[H\Sigma_t H^T + R]^{-1}(y_t - H\mu_t)\right].}$$

2.4 GP Regression as Kalman Smoothing

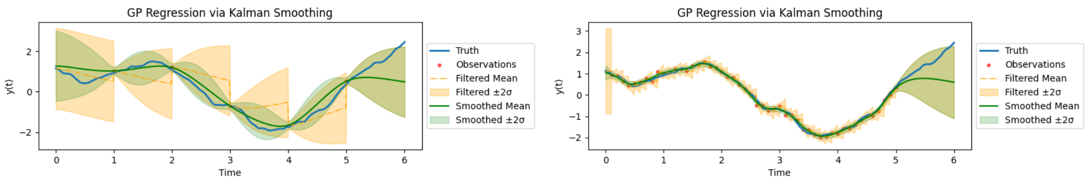

Figure 1: Gaussian Process Regression via Kalman Smoothing

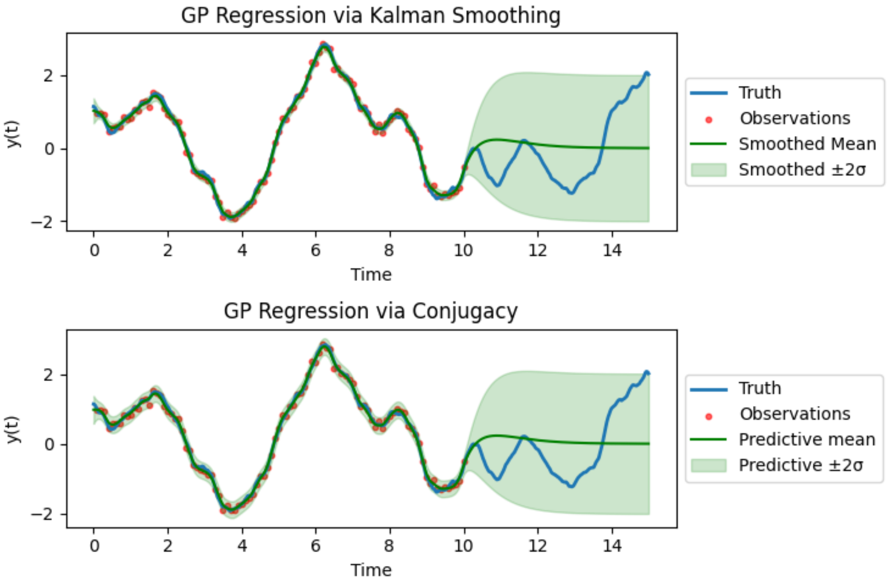

Figure 2: Gaussian Process Regression Comparison: Kalman Smoothing v.s. Conjugacy

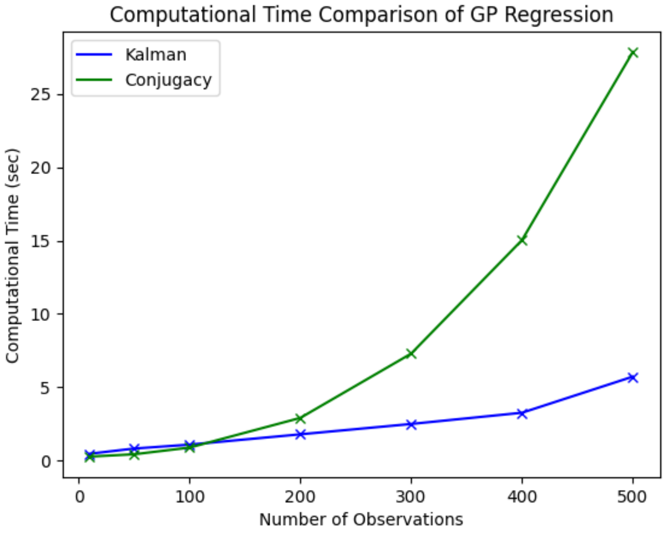

Figure 3: Computational Time of Gaussian Process Regression Comparison: Kalman Smoothing v.s. Conjugacy