Probabilistic Inference & Learning

Exercise Sheet #8

Gauss-Markov Models

1. **Kalman Filters** In Lecture 9, it was shown that inference in Markov chains can be performed through a linear-cost filtering and smoothing algorithm. The goal of this exercise is to show what was only noted without proof in the lecture: that for linear time-invariant Gaussian models, this algorithm reduces to the famous *Kalman filter*.
   Consider a time series of latent states $X := [x_0, x_1, \dots, x_n] \in \mathbb{R}^{d\times(n+1)}$. Assume that their joint distribution $p(X)$ has the Markov property $p(x_t \mid x_{0:t-1}) = p(x_t \mid x_{t-1})\, \forall t = 1, \dots, n$ (the notation $X_{a:b} = [x_a, x_{a+1}, \dots, x_b]$ denotes a slice/subset of a vector). Further, consider observations $Y := [y_0, \dots, y_n] \in \mathbb{R}^{m\times(n+1)}$ with the factorizing likelihood $p(y_t \mid X) = p(y_t \mid x_t)$. In the lecture it was shown that this structure implies that the *predictive*, *estimation* and *smoothed* distributions $p(x_t \mid Y_{0:t-1})$, $p(x_t \mid Y_{0:t})$ and $p(x_t \mid Y)$, respectively, can be computed recursively by the so-called *Chapman-Kolmogorov Equation* (CK), Bayes' Theorem (B), and the *smoothing* equation (S):

$$p(x_t \mid Y_{0:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1} \tag{CK}$$

$$p(x_t \mid Y_{0:t}) = \frac{p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})}{p(y_t)}. \tag{B}$$

$$p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \frac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{0:t})}\, dx_{t+1} \tag{S}$$

   Now assume that the conditional distributions defining the generative model for X, Y are given by the following *linear time-invariant (LTI) Gaussian* model:

$$p(x_0) = \mathcal{N}(x_0; m_0^-, P_0^-) \qquad p(x_t \mid x_{t-1}) = \mathcal{N}(x_t; Ax_{t-1}, Q) \qquad p(y_t \mid x_t) = \mathcal{N}(y_t; Hx_t, R), \tag{1}$$

   with matrices and vectors (spd. = symmetric positive definite)

$$m_0^- \in \mathbb{R}^d, \quad P_0^-, Q \in \mathbb{R}^{d\times d} \text{ spd.}, \quad A \in \mathbb{R}^{d\times d}, \quad H \in \mathbb{R}^{m\times d}, \quad R \in \mathbb{R}^{m\times m} \text{ spd.} \tag{2}$$

   Using the properties of Gaussian distributions introduced in lecture 3 (slide 17), show that under these assumptions, Equations (CK), (B) and (S) take on the explicit forms

$$\begin{aligned}
p(x_t \mid Y_{0:t-1}) &= \mathcal{N}(x_t; m_t^-, P_t^-) &&= \mathcal{N}(x_t; Am_{t-1}, AP_{t-1}A^\intercal + Q) && \text{(Kalman prediction)}\\
p(x_t \mid Y_{0:t}) &= \mathcal{N}(x_t; m_t, P_t) &&= \mathcal{N}(x_t; m_t^- + K_t(y_t - Hm_t^-), (I - KH)P_t^-) && \text{(Kalman estimation)}\\
p(x_t \mid Y) &= \mathcal{N}(x_t; m_t^s, P_t^s) &&= \mathcal{N}(x_t; m_t + G_t(m_{t+1}^s - m_{t+1}^-), P_t + G_t(P_{t+1}^s - P_{t+1}^-)G_t^\intercal)) && \text{(Rauch-Tung-Striebel smoothed estimation)}
\end{aligned}$$

   with the *Kalman gain* K_t and *smoother gain* G_t given by, respectively

$$K_t := P_t^- H^\intercal (HP_t^- H^\intercal + R)^{-1} \qquad \text{and} \qquad G_t := P_t A^\intercal (P_{t+1}^-)^{-1}. \tag{3}$$

   50 points

2. **Stochastic Differential Equations** In Lecture 13, the LTI stochastic differential equation

$$dx(t) = Fx\, dt + L\, d\omega \qquad \text{with} \qquad x(t_0) = x_0, F \in \mathbb{R}^{d\times d}, L \in \mathbb{R}^d \tag{4}$$

   was defined as a reformulation of the Gaussian process over the function $x : t \in \mathbb{R} \mapsto x(t) \in \mathbb{R}^d$ with mean and covariance (i.e. kernel) function

$$m(t) = e^{F(t-t_0)} x_0 \qquad k(t_a, t_b) = \int_{t_0}^{\min(t_a,t_b)} e^{F(t_a-\tau)} L L^\intercal e^{F^\intercal (t_b-\tau)}\, d\tau \tag{5}$$

where e^X is the matrix exponential function

$$e^X := \sum_{n=0}^{\infty} \frac{X^n}{n!} \qquad \text{and} \qquad X^n := \underbrace{X \cdot X \cdots X}_{n \text{ times}} \tag{6}$$

Consider the choices

$$(a) \qquad F = 0 \qquad L = \theta \qquad \textit{((10 points))}$$

$$(b) \qquad F = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad L = \begin{bmatrix} 0 \\ \theta \end{bmatrix} \qquad \textit{((15 points))}$$

$$(c) \qquad F = -\xi \qquad L = \theta \qquad \textit{((10 points))}$$

$$(d) \qquad F = \begin{bmatrix} 0 & 1 \\ -\xi^2 & -2\xi \end{bmatrix} \qquad L = \begin{bmatrix} 0 \\ \theta \end{bmatrix}. \qquad \textit{((15 points))}$$

These four choices identify the Wiener process (a), the integrated Wiener process (b) (whose posterior mean is associated with cubic spline interpolants), and the first two members (c,d) of the so-called *Matérn-family* of kernels.

Use Eq. (5) to find explicit forms for the mean function $m(t)$ and kernel $k(t_a, t_b)$.

Some hints:
For (b), note that $F^2 = 0$ (and thus $F^k = 0 \; \forall k \geq 2$ – the matrix is *nilpotent*). This property can be used in Eq. (6) to compute $\exp(Ft)$.
For (d), note that if a matrix X has eigenvalue decomposition $X = VDV^{-1}$, then its matrix exponential can be written as $e^X = Ve^D V^{-1}$, where e^D is a diagonal matrix containing the (scalar) exponentials of the eigenvalues of X. So you can just compute the eigenvalue decomposition of this 2×2 matrix F to find $\exp F$. This can be done manually for such small matrices (linear algebra reminder: just solve the characteristic polynomial $\det(\lambda I - F)$ for the eigenvalues λ, then find the eigenvectors by solving $(F - \lambda I)v = 0$ for v).

Background Information (only for those interested, not necessary to solve the exercise)

The matrix F in (d) is what is known as a *companion matrix*: It is of the form

$$F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -c_0 & -c_1 & -c_2 & \cdots & -c_{q-1} \end{bmatrix}$$

Such a matrix is the "companion" of the polynomial

$$p(t) = c_0 + c_1 t + \cdots + c_{q-1} t^{q-1} + t^q,$$

because the characteristic polynomial of F is equal to p. The Matérn family (for $q \in \mathbb{N}$)

$$k_{q+1/2}(r := |t_a - t_b|) = \theta^2 \frac{\Gamma(q+1)}{\Gamma(2q+1)} \sum_{i=0}^{q} \frac{(q+i)!}{i!(q-i)!} \left(\sqrt{8\nu}\frac{r}{\lambda}\right)^{q-i} \cdot \exp\left(-\sqrt{2\nu}\frac{r}{\lambda}\right).$$

is the family of kernels associated with SDEs of the form as in (c), (d), with F the companion matrix of the polynomial $(\xi + it)^{-(q+1)}$ with $\xi = \sqrt{2\nu}\lambda$. Thus, F only has a single, degenerate, eigenvalue $\lambda = -\xi$. That is, they correspond to state space models $x(t) = [x_0, x_1, \ldots, x_q]$ where the state consists of the first q derivatives of x_0 (due to the 1's in F). Thus, the Matérn class provides the basic GP model for q-times continuously differentiable and stationary functions. The base case $q = 0$ (c) is known as the *Ornstein-Uhlenbeck (OU)* process. Similar to how the Wiener process models Brownian motion of free particles in an ideal gas, the OU process models the velocity of such particles when bound by a harmonic potential.