Exercises
Probabilistic Artificial Intelligence
Fall 2023

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause

Homework #No. 2 (Gaussian Processes)

For questions, please refer to Moodle.
Released on **12/10/2023**

General Instructions

- Submission of solutions is not mandatory but solving the exercises are highly recommended. The master solution will be released after the exercise deadline.
- Part of the exercises are available on Moodle as a quiz. These problems are marked with [☑].

Exercise 1: Gaussian Process Kernels

Given a Gaussian process $GP(\mu, k)$ indexed by $\mathbb{R}$, which is the set of all real numbers, with mean function $\mu : \mathbb{R} \to \mathbb{R}$ and kernel function $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. In the following, we assume that we have a RBF kernel function with lengthscale and variance of 1, i.e. $k(x, x') = e^{-\frac{(x-x')^2}{2}}$, and a linear mean function, i.e. $\mu(x) = x$. We want to model an unknown function $f : \mathcal{X} \to \mathbb{R}$.

(a) [☑]Given three data points $x_1 = 1, x_2 = 3, x_3 = 9$, what are the mean vector $\mathbf{m}$ and covariance matrix K of the marginal distribution $(f(x_1), f(x_2), f(x_3)) \sim \mathcal{N}(\mathbf{m}, K)$?

√ $\mathbf{m} = \begin{bmatrix} 1 & 3 & 9 \end{bmatrix}, K = \begin{bmatrix} 1 & e^{-2} & e^{-32} \\ e^{-2} & 1 & e^{-18} \\ e^{-32} & e^{-18} & 1 \end{bmatrix}$

○ $\mathbf{m} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, K = \begin{bmatrix} 1 & e^{-2} & e^{-32} \\ e^{-2} & 1 & e^{-18} \\ e^{-32} & e^{-18} & 1 \end{bmatrix}$

○ $\mathbf{m} = \begin{bmatrix} e^{-1} & e^{-3} & e^{-9} \end{bmatrix}, K = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

○ $\mathbf{m} = \begin{bmatrix} e^{-1} & e^{-3} & e^{-9} \end{bmatrix}, K = \begin{bmatrix} 0 & 2 & 8 \\ 2 & 0 & 6 \\ 8 & 6 & 0 \end{bmatrix}$

Solution: Gaussian distributions are closed under marginalization, meaning the probability distribution of $(f(x_1), f(x_2), f(x_3))$ is Gaussian with:

mean vector $\mathbf{m} = \begin{bmatrix} \mu(x_1), \mu(x_2), \mu(x_3) \end{bmatrix}$ and covariance matrix $K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & k(x_1, x_3) \\ k(x_2, x_1) & k(x_2, x_2) & k(x_2, x_3) \\ k(x_3, x_1) & k(x_3, x_2) & k(x_3, x_3) \end{bmatrix}$

By plugging in $\mu(x) = x$ and $k(x, x') = e^{-\frac{(x-x')^2}{2}}$, we obtain :

$$\mathbf{m} = \begin{bmatrix} 1, 3, 9 \end{bmatrix}, K = \begin{bmatrix} 1 & e^{-2} & e^{-32} \\ e^{-2} & 1 & e^{-18} \\ e^{-32} & e^{-18} & 1 \end{bmatrix}$$

(b) [☑]We are now given noise-free observation for $f(x_1) = 3$ and $f(x_2) = 10$. What is the mean m_3^p and variance σ_3^p of of the posterior distribution $f(x_3) | f(x_1), f(x_2) \sim N(m_3^p, \sigma_3^2)$?

○ $m_3^p = 9 + \frac{2(e^{-32} - e^{-20}) - 7(e^{-34} - e^{-18})}{1 - e^{-4}}$, $\sigma_3^2 = 1 - e^{-100} \frac{1 - e^{-2}}{1 - e^{-4}}$

○ $m_3^p = 9$, $\sigma_3^2 = 1 - e^{-100} \frac{1 - e^{-2}}{1 - e^{-4}}$

○ $m_3^p = 9$, $\sigma_3^2 = 1 - \frac{e^{-64} + e^{-36} - 2e^{-52}}{1 - e^{-4}}$

√ $m_3^p = 9 + \frac{2(e^{-32} - e^{-20}) - 7(e^{-34} - e^{-18})}{1 - e^{-4}}$, $\sigma_3^p = 1 - \frac{e^{-64} + e^{-36} - 2e^{-52}}{1 - e^{-4}}$

Solution: Again, we know that Gaussian distributions are closed under conditioning. If we define, $\mathbf{f} = [f(x_1), f(x_2), f(x_3)]$, $A = \{1,2\}$ and $B = \{3\}$, then we can re-write the followings:

$$\begin{aligned}
\mathbf{f} &= \begin{bmatrix} \mathbf{f}_A \\ \mathbf{f}_B \end{bmatrix} \\
\mathbf{m} &= \begin{bmatrix} \mathbf{m}_A \\ \mathbf{m}_B \end{bmatrix} \\
K &= \begin{bmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{bmatrix}
\end{aligned}$$

Then from conditioning on A, we have:

$$\begin{aligned}
m_3^p &= \mathbf{m}_{B|A} = \mathbf{m}_B + K_{BA}K_{AA}^{-1}(\mathbf{f}_A - \mathbf{m}_A) \\
\sigma_3^2 &= K_{B|A} = K_{BB} - K_{BA}K_{AA}^{-1}K_{AB}
\end{aligned}$$

From the previous questions, all quantities are known except K_{AA}^{-1}. Luckily, K_{AA} is a 2×2 matrix, hence inverting it is straight forward.

$$K_{AA}^{-1} = \frac{1}{k_{11}k_{22} - k_{21}k_{12}} \begin{bmatrix} k_{22} & -k_{12} \\ -k_{21} & k_{11} \end{bmatrix} = \frac{1}{D} \begin{bmatrix} k_{11} & -k_{12} \\ -k_{12} & k_{11} \end{bmatrix}$$

with $D = \frac{1}{k_{11}^2 - k_{12}^2}$ **Computing the mean**

$$\begin{aligned}
K_{BA}K_{AA}^{-1} &= \frac{1}{D} \begin{bmatrix} k_{31} & k_{32} \end{bmatrix} \begin{bmatrix} k_{22} & -k_{12} \\ -k_{21} & k_{11} \end{bmatrix} \\
&= \frac{1}{D} \begin{bmatrix} k_{31}k_{22} - k_{21}k_{32} & -k_{31}k_{12} + k_{11}k_{32} \end{bmatrix} \\
&= \frac{1}{D} \begin{bmatrix} k_{31}k_{11} - k_{12}k_{32} & -k_{31}k_{12} + k_{11}k_{32} \end{bmatrix}
\end{aligned}$$

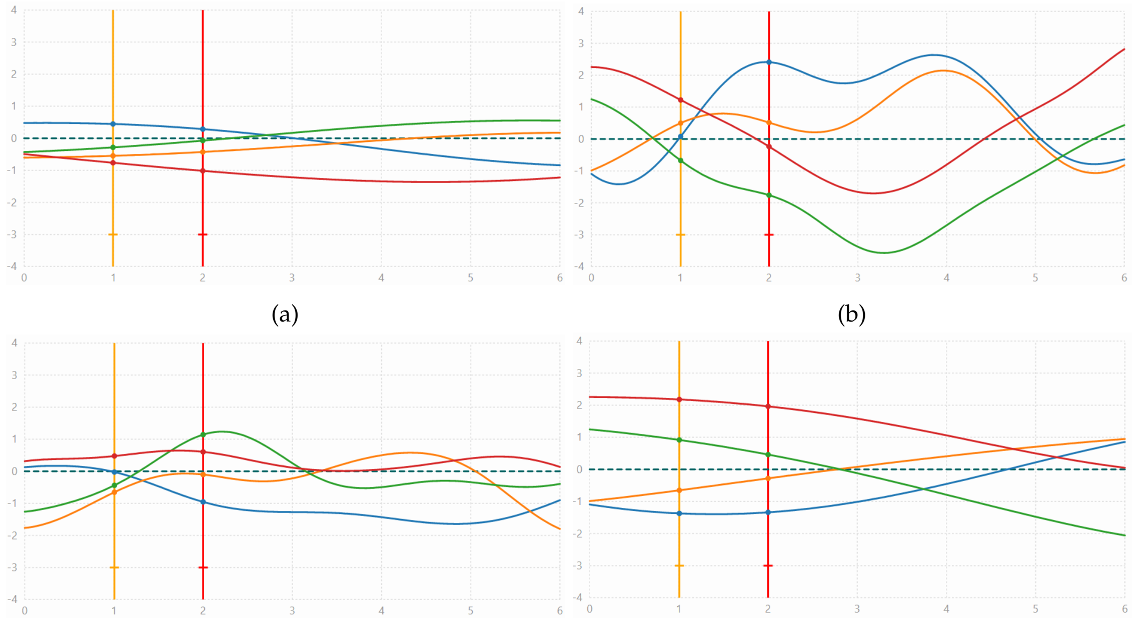
$$(\mathbf{f}_A - \mathbf{m}_A) = \begin{bmatrix} f(x_1) - x_1 \\ f(x_2) - x_2 \end{bmatrix}$$

Hence

$$\begin{aligned}
m_3^p &= x_3 + \frac{(k_{31}k_{11} - k_{12}k_{32})(f(x_1) - x_1) - (k_{31}k_{12} - k_{11}k_{32})(f(x_2) - x_2)}{D} \\
&= 9 + \frac{2(e^{-32} - e^{-20}) - 7(e^{-34} - e^{-18})}{1 - e^{-4}}
\end{aligned}$$

Computing the variance Using above calculation, we have:

$$\begin{aligned}
K_{BA}K_{AA}^{-1}K_{AB} &= \frac{1}{D} \begin{bmatrix} k_{31}k_{11} - k_{12}k_{32} & -k_{31}k_{12} + k_{11}k_{32} \end{bmatrix} \begin{bmatrix} k_{13} \\ k_{23} \end{bmatrix} \\
&= \frac{k_{13}(k_{31}k_{11} - k_{12}k_{32}) + k_{23}(k_{11}k_{32} - k_{31}k_{12})}{D} \\
&= \frac{k_{13}^2 k_{11} - 2k_{12}k_{23}k_{13} + k_{11}k_{23}^2}{D} \\
&= \frac{e^{-64} + e^{-36} - 2e^{-52}}{1 - e^{-4}}
\end{aligned}$$



(a)

(b)

(c)

(d)

Figure 1: Samples drawn from Gaussian process with mean 0 and RBF kernel with four pairs of parameters. The x-axis is t and the y-axis is X_t. Each plot is generated using a specific pair of variance (σ^2) and length scale (l) parameters. Different colors represent different sample functions drawn from a Gaussian process. Function values at $t = 1$ (i.e., X_1) and $t = 2$ (i.e., X_2) are plotted as colored dots for simpler comparison.

Hence :

$$\sigma_3^2 = 1 - \frac{e^{-64} + e^{-36} - 2e^{-52}}{1 - e^{-4}}$$

(c) [☑]We now investigate the influence of kernel parameters on the prior distribution of functions. Imagine that you draw sample functions from Gaussian process with mean 0 and RBF kernel with four pairs of variance and length scale parameters.

$$k(t, t') = \sigma^2 \exp\left(-\frac{(t - t')^2}{2l^2}\right)$$

1. $\sigma^2 = 0.5, l = 1$.
2. $\sigma^2 = 0.5, l = 4$.
3. $\sigma^2 = 2, l = 1$.
4. $\sigma^2 = 2, l = 4$.

Samples are plotted in Figure 1 in no particular order, each plot for samples drawn from **one** particular pair of parameters.

Applying the results from the previous question, which plot **most likely** corresponds to which pair of parameters?

1. $\sigma^2 = 0.5, l = 1$ ○ Fig.1a ○ Fig.1b √ **Fig.1c** ○ Fig.1d
2. $\sigma^2 = 0.5, l = 4$ √ **Fig.1a** ○ Fig.1b ○ Fig.1c ○ Fig.1d
3. $\sigma^2 = 2, l = 1$ ○ Fig.1a √ **Fig.1b** ○ Fig.1c ○ Fig.1d
4. $\sigma^2 = 2, l = 4$ ○ Fig.1a ○ Fig.1b ○ Fig.1c √ **Fig.1d**

Solution: 1c, 2a, 3b, 4d. For RBF kernel, σ^2 corresponds to the vertical scaling of functions. Thus, by comparing the magnitude of the fluctuations of the sample functions in different plots, we can conclude that $\sigma^2 = 0.5$ corresponds to (a) and (c), while $\sigma^2 = 2$ corresponds to (b) and (d). From the derivation in the previous question, we can conclude that for any fixed $\sigma > 0$, as l increases, $\mathbb{V}ar\left(\frac{X_1 - X_2}{\sigma}\right)$ decreases and thus $P(|X_1 - X_2| \leq \sigma)$ increases. In other words, $|X_1 - X_2|$ is most likely to be bounded by σ.

Exercise 2: Gaussian Processes Regression With Linear Kernel

[✓]*Gaussian process (GP)*, denoted as $GP(\mu, k)$, is a stochastic process specified by some mean function $\mu : \mathcal{X} \to \mathbb{R}$ and some kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In this exercise, you will show that Bayesian linear regression yields the same prediction as Gaussian process regression with the linear kernel $k(x, x') = \lambda x^T x'$.

Consider an unknown function $f : \mathcal{X} \to \mathbb{R}$ and a dataset $A = \{(x_1, y_1), \dots, (x_m, y_m)\}$ of noise-perturbed evaluations $y_i = f(x_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$. Now, our task is to predict the distribution of f in a new point $x^* \in \mathcal{X}$.

GP regression: A priori, we assume $f \sim GP(0, k)$ with linear kernel $k(x, x') = \lambda x^T x'$ (since we want to emulate BLR). In this case, the posterior update for $\mu'(x)$ and $k'(x, x')$ based on the evaluations $y_A = [y_1, \dots, y_m]^T$ can be computed as follows:

$$\begin{aligned} \mu'(x) &= \mu(x) + k_{x,A}^T (K_{A,A} + \sigma_n^2 I)^{-1} y_A, \\ k'(x, x') &= k(x, x') - k_{x,A}^T (K_{A,A} + \sigma_n^2 I)^{-1} k_{x,A}, \end{aligned}$$

where $K_{A,A} \in \mathbb{R}^{m \times m}$ is a matrix with elements $[K_{A,A}]_{i,j} = k(x_i, x_j)$, and $k_{x,A} \in \mathbb{R}^m$ is a vector with elements $[k_{x,A}]_i = k(x, x_i)$.

BLR: In the Bayesian Linear regression, we assume the liner model $f(x_i) = x_i^T w$ with the prior over weights $p(w) = \mathcal{N}(0, \sigma_p^2)$. In the class, we have shown that for the evaluations y_A (or, $y_{1:n}$ in the lecture notation):

$$\begin{aligned} p(w \mid y_A) &= \mathcal{N}(\bar{\mu}, \bar{\Sigma}), \\ \bar{\mu} &= \frac{1}{\sigma_n^2} \bar{\Sigma} X^T y_A, \\ \bar{\Sigma} &= \left(\frac{1}{\sigma_n^2} X^T X + \frac{1}{\sigma_p^2} I \right)^{-1}, \end{aligned}$$

where $X \in \mathbb{R}^{m \times d}$ consists of rows $x_1^T, x_2^T, \dots, x_m^T$.

Given a new point x^*, find the predictive distribution both for BLR and GP regression. For which value λ do they coincide?

Hint: One of the Woodbury identities might be useful.

○ $\lambda = \frac{1}{\sigma_n^2} + \frac{1}{\sigma_p^2}$ √ $\lambda = \sigma_p^2$ ○ $\lambda = \sigma_n^2$ ○ $\lambda = \frac{1}{\sigma_p^2}$

Solution: Following the posterior BLR and $f(x^*) = {x^*}^T w$, then:

$$f(x^*) \sim \mathcal{N}({x^*}^T \bar{\mu}, {x^*}^T \bar{\Sigma} x^*) = \mathcal{N}\left({x^*}^T \frac{1}{\sigma_n^2} \bar{\Sigma} X^T y_A, {x^*}^T \bar{\Sigma} x^* \right).$$

Let us now look at the GP model: $f(x^*) \sim \mathcal{N}(\mu'(x^*), k'(x^*, x^*))$.

$$\mu'(x^*) = \mu(x^*) + k_{x^*,A}^T (K_{A,A} + \sigma_n^2 I)^{-1} y_A$$
$$k'(x^*, x^*) = k(x^*, x^*) - k_{x^*,A}^T (K_{A,A} + \sigma_n^2 I)^{-1} k_{x^*,A}.$$

Now we write $K_{A,A}$ and $k_{x^*,A}$ in a more convenient way:

$$K_{A,A} = \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ k(x_m, x_1) & \dots & k(x_m, x_m) \end{bmatrix} = \lambda \begin{bmatrix} x_1^T x_1 & \dots & x_1^T x_m \\ \vdots & \ddots & \vdots \\ x_m^T x_1 & \dots & x_m^T x_m \end{bmatrix} = \lambda X X^T,$$

$$k_{x^*,A} = \begin{bmatrix} k(x_1, x^*) \\ \vdots \\ k(x_m, x^*) \end{bmatrix} = \lambda \begin{bmatrix} x_1^T x^* \\ \vdots \\ x_m^T x^* \end{bmatrix} = \lambda X x^*.$$

Then, for the mean we have:

$$\begin{aligned}
\mu'(x^*) &= \overbrace{\mu(x^*)}^{0} + k_{x^*,A}^T (K_{A,A} + \sigma_n^2 I)^{-1} y_A \\
&= (\lambda X x^*)^T (\lambda X X^T + \sigma_n^2 I)^{-1} y_A \\
&= \frac{\lambda}{\sigma_n^2} {x^*}^T X^T \left(\frac{\lambda X}{\sigma_n^2} X^T + I \right)^{-1} y_A \\
&\overset{(i)}{=} \frac{\lambda}{\sigma_n^2} {x^*}^T \left(X^T \frac{\lambda X}{\sigma_n^2} + I \right)^{-1} X^T y_A \\
&= \frac{1}{\sigma_n^2} {x^*}^T \left(\frac{1}{\sigma_n^2} X^T X + \frac{1}{\lambda} I \right)^{-1} X^T y_A \\
&\overset{(ii)}{=} {x^*}^T \frac{1}{\sigma_n^2} \bar{\Sigma} X^T y_A, \qquad \text{yielding the same prediction as BLR.}
\end{aligned}$$

Where in (i) we use the Woodbury push-through identity, i.e., $U(VU + I)^{-1} = (UV + I)^{-1} U$, and in (ii) we take $\lambda = \sigma_p^2$.
Now we look at the variance in the GP model, which is $k'(x^*, x^*)$.

$$\begin{aligned}
k'(x^*, x^*) &= \sigma_p^2 {x^*}^T x^* - (\sigma_p^2)^2 {x^*}^T X^T (\sigma_p^2 X X^T + \sigma_n^2 I)^{-1} X x^* \\
&= \sigma_p^2 {x^*}^T (I - \sigma_p^2 X^T (\sigma_p^2 X X^T + \sigma_n^2 I)^{-1} X) x^* \\
&= \sigma_p^2 {x^*}^T \left(I - \frac{\sigma_p^2}{\sigma_n^2} X^T \left(I + \frac{\sigma_p^2}{\sigma_n^2} X X^T \right)^{-1} X \right) x^* \\
&\overset{(i)}{=} \sigma_p^2 {x^*}^T \left(I - \frac{\sigma_p^2}{\sigma_n^2} X^T X \left(I + \frac{\sigma_p^2}{\sigma_n^2} X^T X \right)^{-1} \right) x^* \\
&= \sigma_p^2 {x^*}^T \left(I + \frac{\sigma_p^2}{\sigma_n^2} X^T X \right)^{-1} x^* \\
&= {x^*}^T \left(\frac{1}{\sigma_p^2} I + \frac{1}{\sigma_n^2} X^T X \right)^{-1} x^* = {x^*}^T \bar{\Sigma} x^*,
\end{aligned}$$

yielding the same predictive variance.

Exercise 3: Kalman filters

Consider the following dynamical system describing a moving particle:

$$X_{t+1} = X_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_x^2),$$

where $X_t \in \mathbb{R}$ denotes a position of the particle at time $t > 0$ and ε_t is a time-independent and identically (i.i.d) distributed Gaussian noise. We would like to track the position of the particle over time, however, this is not easy. We cannot directly observe X_t but only see its measurement Y_t perturbed by i.i.d Gaussian noise η_t:

$$Y_t = X_t + \eta_t \qquad \eta_t \sim \mathcal{N}(0, \sigma_y^2).$$

To study this process, we divide the problem in prediction and conditioning.

(a) [✓]*Prediction*: assume the posterior distribution of X_t is a Gaussian with mean μ_t and variance σ_t^2 after having observed $y_{1:t} := \{Y_1 = y_1, \ldots, Y_t = y_t\}$:

$$p(X_t \mid y_{1:t}) = \mathcal{N}(\mu_t, \sigma_t^2).$$

What is the predictive distribution for the particle's position at the next time step $p(X_{t+1} \mid y_{1:t})$:

○ $\mathcal{N}(0, \sigma_t^2 + t\sigma_y^2)$ ○ $\mathcal{N}(\mu_t, \sigma_y^2 + \sigma_x^2)$ √ $\mathcal{N}(\mu_t, \sigma_t^2 + \sigma_x^2)$ ○ $\mathcal{N}(\mu_t, \sigma_t^2 + \sigma_y^2)$

Solution: Since X_t and ε_t are normally distributed, $p(X_{t+1} \mid y_{1:t}) = p(X_t + \varepsilon_t \mid y_{1:t})$ is also Gaussian (being a sum of two independent Gaussian variables). Then

- $\mathbb{E}[X_{t+1} \mid y_{1:t}] = \mathbb{E}[X_t \mid y_{1:t}] + \overbrace{\mathbb{E}[\varepsilon_t]}^{=0} = \mu_t$
- $\mathbb{V}ar(X_{t+1} \mid y_{1:t}) = \mathbb{V}ar(X_t \mid y_{1:t}) + \mathbb{V}ar(\varepsilon_t) = \sigma_t^2 + \sigma_x^2$

Therefore $p(X_{t+1} \mid y_{1:t}) = \mathcal{N}(\mu_t, \sigma_t^2 + \sigma_x^2)$.

(b) [✓]*Conditioning*: once we have the prediction distribution $p(X_{t+1} \mid y_{1:t})$, we would like to understand how acquiring y_{t+1} would help. For what value k_{t+1}, called the Kalman gain, can we write $p(X_{t+1} \mid y_{1:t}, y_{t+1})$ as follows:

$$\begin{aligned} p(X_{t+1} \mid y_{1:t}, y_{t+1}) &= \mathcal{N}(\mu_{t+1}, \sigma_{t+1}^2), \\ \text{with} \quad \mu_{t+1} &= \mu_t + k_{t+1}(y_{t+1} - \mu_t) \\ \text{and} \quad \sigma_{t+1}^2 &= (1 - k_{t+1})(\sigma_x^2 + \sigma_t^2) \end{aligned}$$

○ $k_{t+1} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2}$ √ $k_{t+1} = \frac{\sigma_x^2 + \sigma_t^2}{\sigma_x^2 + \sigma_t^2 + \sigma_y^2}$ ○ $k_{t+1} = \frac{\sigma_x^2 + \sigma_t^2 + \sigma_y^2}{\sigma_x^2 + \sigma_t^2}$ ○ $k_{t+1} = \frac{\sigma_x^2 + \sigma_t^2}{\sigma_y^2}$

Solution: Let's start with rewriting the joint distribution as:

$$\begin{bmatrix} X_{t+1} \mid y_{1:t} \\ Y_{t+1} \mid y_{1:t} \end{bmatrix} = \begin{bmatrix} X_{t+1} \mid y_{1:t} \\ X_{t+1} \mid y_{1:t} + \eta_{t+1} \mid y_{1:t} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} X_{t+1} \mid y_{1:t} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \eta_{t+1}.$$

Since $\begin{bmatrix} X_{t+1} \mid y_{1:t} \\ Y_{t+1} \mid y_{1:t} \end{bmatrix}$ is a linear combination of the Gaussians $X_{t+1} \mid y_{1:t}$ and η_{t+1}, the joint distribution is also Gaussian and given by:

$$\begin{aligned} \begin{bmatrix} X_{t+1} \mid y_{1:t} \\ Y_{t+1} \mid y_{1:t} \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} \mu_t, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mathbb{V}ar(X_{t+1} \mid y_{1:t}) \begin{bmatrix} 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \mathbb{V}ar(\eta_{t+1}) \begin{bmatrix} 0 & 1 \end{bmatrix} \right) \\ &= \mathcal{N}\left(\begin{bmatrix} \mu_t \\ \mu_t \end{bmatrix}, \begin{bmatrix} \sigma_t^2 + \sigma_x^2 & \sigma_t^2 + \sigma_x^2 \\ \sigma_t^2 + \sigma_x^2 & \sigma_t^2 + \sigma_x^2 + \sigma_y^2 \end{bmatrix} \right). \end{aligned}$$

Finally, given the above joint distribution over $X_{t+1} \mid y_{1:t}$ and $Y_{t+1} \mid y_{1:t}$, we can condition on the latter and use the formulas for multivariate Gaussian conditional distributions to get the following closed form:

$$p(X_{t+1} \mid Y_{t+1} = y_{t+1}, y_{1:t}) = \mathcal{N}(\mu_{t+1}, \sigma^2_{t+1}),$$

$$\mu_{t+1} = \mu_t + \underbrace{\frac{\sigma_x^2 + \sigma_t^2}{\sigma_x^2 + \sigma_t^2 + \sigma_y^2}}_{=k_{t+1}} (y_{t+1} - \mu_t),$$

$$\sigma^2_{t+1} = \sigma_x^2 + \sigma_t^2 - \underbrace{\frac{\sigma_x^2 + \sigma_t^2}{\sigma_x^2 + \sigma_t^2 + \sigma_y^2}}_{=k_{t+1}} (\sigma_x^2 + \sigma_t^2),$$

where the substitution of k_{t+1} would lead to the required result.

We now would like to discuss the importance of the Kalman gain k_t. This value plays an important role in the interpretation of the conditioning step.

(c) How the different parameters σ_x, σ_y and σ_t influence the value of the Kalman gain?

Solution: First of all $k_t \in [0, 1]$.

- σ_t grows $\Rightarrow k_t \to 1$
- σ_x grows $\Rightarrow k_t \to 1$
- σ_y grows $\Rightarrow k_t \to 0$

(d) How the value of the Kalman gain k_{t+1} gives more importance to y_{t+1} or, gives more importance to the previous knowledge μ_{t+1}? Can you guess why it's called "gain"?

Solution: If $k_t = 1$, then we have that $\mu_t = y_t$. If $k_t = 0$, $\mu_t = \mu_{t-1}$. In general, if k_t is close to 0, we give more importance to previous knowledge. If k_t is close to 1, we give more importance to y_t, which is the new observation.

The Kalman gain, then, is a measure of how much information we *gain* from the new observation.

(e) [☑]Now we will show that the Kalman filter can be seen as a GP. To this end, we define:

$$f : \mathbb{Z}_+ \to \mathbb{R} \quad \text{such that } f(t) = X_t.$$

Assuming that $X_0 \sim \mathcal{N}(0, \sigma_0^2)$ and $X_{t+1} = X_t + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0, \sigma_x^2)$, we can show that $f \sim \mathcal{GP}(0, k_{KF})$ with which kernel function k_{KF}

○ $k_{KF}(t, t') = \sigma_x^2 + \sigma_0^2 \max\{t, t'\}$ ○ $k_{KF}(t, t') = \sigma_x^2 + \sigma_0^2 \min\{t, t'\}$
√ $k_{KF}(t, t') = \sigma_0^2 + \sigma_x^2 \min\{t, t'\}$ ○ $k_{KF}(t, t') = \sigma_0^2 + \sigma_x^2 \max\{t, t'\}$

Solution: First we look at the mean:

$$\mu(t) = \mathbb{E}[X_t] = \mathbb{E}[X_{t-1} + \varepsilon_{t-1}] = \mathbb{E}[X_{t-1}] = \mu(t-1).$$

Knowing that $\mu(0) = 0$, we can derive that $\mu(t) = 0,\ \forall t \in \mathbb{Z}_+$. Now we look at the variance of X_t:

$$\mathbb{V}ar(X_t) = \mathbb{V}ar(X_0 + \varepsilon_0 + \cdots + \varepsilon_{t-1}) = \sigma_0^2 + t\sigma_x^2.$$

Now, to finish, we look at the distribution of $[f(t), f(t')]^T$, assuming $t < t'$.

$$\begin{bmatrix} X_t \\ X_{t'} \end{bmatrix} = \begin{bmatrix} X_t \\ X_t \end{bmatrix} + \begin{bmatrix} 0 \\ \varepsilon_t \end{bmatrix} + \cdots + \begin{bmatrix} 0 \\ \varepsilon_{t'-1} \end{bmatrix}.$$

Therefore we get that

$$\begin{aligned} \begin{bmatrix} X_t \\ X_{t'} \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbb{V}ar(X_t) \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + (t'-t) \begin{bmatrix} 0 & 0 \\ 0 & \sigma_x^2 \end{bmatrix} \right) \\ &= \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbb{V}ar(X_t) & \mathbb{V}ar(X_t) \\ \mathbb{V}ar(X_t) & \mathbb{V}ar(X_t) + (t'-t)\sigma_x^2 \end{bmatrix} \right). \end{aligned}$$

Finally, we take the kernel $k_{KF}(t, t')$ to be the covariance between $f(t)$ and $f(t')$, which is $\mathbb{V}ar(X_t) = \sigma_0^2 + \sigma_x^2 t$. Notice however, that we chose $t \leq t'$. Otherwise, if $t > t'$ we get the opposite. Overall the kernel is:

$$k_{KF}(t, t') = \sigma_0^2 + \sigma_x^2 \min\{t, t'\}.$$

Remark: Note that this particular kernel $k(t, t') = \min\{t, t'\}$ (but over the continuous-time domain) defines the stochastic Wiener process (also known as Brownian motion).

Exercise 4: Hyperparameters Selection and Marginal Likelihood

Consider an unknown function $f : \mathcal{X} \to \mathbb{R}$ and a dataset $A = \{X, \mathbf{y}\} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of noise-perturbed evaluations $y_i = f(x_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$, we make the hypothesis that $f \sim GP(0, k_\theta)$, with zero mean function and covariance function $k_\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We are interested into selecting hyperparameters θ by maximizing the marginal likelihood $p(\mathbf{y} \mid X, \theta)$.

(a) [☑]We define $\mathbf{K}_{\mathbf{y},\theta} = \mathbf{K}_{f,\theta} + \sigma_n^2 I$ as the covariance matrix of $\mathbf{y}$ for covariance function k_θ. We also define $\alpha = \mathbf{K}_{\mathbf{y},\theta}^{-1}\mathbf{y}$. What is the value of the the marginal likelihood gradient with respect to θ_j, $\frac{\partial}{\partial \theta_j} \log p(\mathbf{y} \mid X, \theta)$:

○ $\frac{1}{2} \operatorname{tr}\left(\left(\alpha\alpha^T - \mathbf{K}_{\mathbf{f},\theta}^{-1} \right) \frac{\partial \mathbf{K}_{\mathbf{f},\theta}}{\partial \theta_j} \right)$ √ $\frac{1}{2} \operatorname{tr}\left(\left(\alpha\alpha^T - \mathbf{K}_{\mathbf{y},\theta}^{-1} \right) \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j} \right)$

○ $2 \operatorname{tr}\left(\left(\alpha\alpha^T - \mathbf{K}_{\mathbf{y},\theta}^{-1} \right) \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j} \right)$ ○ $2 \operatorname{tr}\left(\left(\alpha\alpha^T - \mathbf{K}_{\mathbf{f},\theta}^{-1} \right) \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j} \right)$

Hint: You can use the following indentites in your derivation:

- For any invertible matrix M, you have:

$$\frac{\partial}{\partial \theta_j} M^{-1} = -M^{-1} \frac{\partial M}{\partial \theta_j} M^{-1}$$

- For any symmetric positive definite matrix S, you have:

$$\frac{\partial}{\partial \theta_j} \log |S| = \operatorname{tr}\left(S^{-1} \frac{\partial S}{\partial \theta_j} \right)$$

Solution: Recall that $y = f(x) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$ and $f \sim \mathcal{N}(0, \mathbf{K}_{f,\theta})$, thus as a sum of Gaussians, we have

$$p(\mathbf{y}|X, \theta) = \mathcal{N}(0, \mathbf{K}_{f,\theta} + \sigma_n^2 I)$$

Thus, we can directly write $\log p(\mathbf{y}|X, \theta)$, as in the lecture:

$$\begin{aligned}\log p(\mathbf{y} \mid X, \theta) &= -\frac{1}{2}\mathbf{y}^\top \left(\mathbf{K}_{f,\theta} + \sigma_n^2 I\right)^{-1} \mathbf{y} - \frac{1}{2} \log \left|\mathbf{K}_{f,\theta} + \sigma_n^2 I\right| - \frac{m}{2} \log 2\pi \\ &= -\frac{1}{2}\mathbf{y}^\top \mathbf{K}_{\mathbf{y},\theta}^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}_{\mathbf{y},\theta}| - \frac{m}{2} \log 2\pi\end{aligned}$$

We can then apply the two indentity provided in the hint to $\frac{\partial \mathbf{K}_{\mathbf{y},\theta}^{-1}}{\partial \theta_j}$ and $\frac{\partial \log |\mathbf{K}_{\mathbf{y},\theta}|}{\partial \theta_j}$, obtaining:

$$\frac{\partial}{\partial \theta_j} \log p(\mathbf{y} \mid X, \boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^\top \mathbf{K}_{\mathbf{y},\theta}^{-1} \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j} \mathbf{K}_{\mathbf{y},\theta}^{-1} \mathbf{y} - \frac{1}{2} \operatorname{tr}\left(\mathbf{K}_{\mathbf{y},\theta}^{-1} \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j}\right)$$

Now using that (1) $\mathbf{y}^\top \mathbf{K}_{\mathbf{y},\theta}^{-1} \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j} \mathbf{K}_{\mathbf{y},\theta}^{-1} \mathbf{y}$ is a scalar, (2) $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, and (3) S^{-1} is symmetric if S is symmetric, we have the following:

$$\begin{aligned}\frac{\partial}{\partial \theta_j} \log p(\mathbf{y} \mid X, \boldsymbol{\theta}) &= \frac{1}{2} \operatorname{tr}\left(\mathbf{y}^\top \mathbf{K}_{\mathbf{y},\theta}^{-1} \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j} \mathbf{K}_{\mathbf{y},\theta}^{-1} \mathbf{y}\right) - \frac{1}{2} \operatorname{tr}\left(\mathbf{K}_{\mathbf{y},\theta}^{-1} \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j}\right) && \text{using (1)} \\ &= \frac{1}{2} \operatorname{tr}\left(\mathbf{K}_{\mathbf{y},\theta}^{-1} \mathbf{y}\mathbf{y}^\top \mathbf{K}_{\mathbf{y},\theta}^{-1} \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j} - \mathbf{K}_{\mathbf{y},\theta}^{-1} \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j}\right) && \text{using (2) and the linearity of trace} \\ &= \frac{1}{2} \operatorname{tr}\left(\mathbf{K}_{\mathbf{y},\theta}^{-1} \mathbf{y} (\mathbf{K}_{\mathbf{y},\theta}^{-1} \mathbf{y})^\top \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j} - \mathbf{K}_{\mathbf{y},\theta}^{-1} \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j}\right) && \text{using (3)} \\ &= \frac{1}{2} \operatorname{tr}\left(\left(\alpha\alpha^T - \mathbf{K}_{\mathbf{y},\theta}^{-1}\right) \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_j}\right)\end{aligned}$$

(b) [☑]We now assume that covariance function for the noisy targets(i.e. including noise contribution) can be written $k_{\mathbf{y}}(x, x') = \theta_0 \tilde{k}(x, x')$ with $\tilde{k}$ a valid kernel independent of θ_0. What is the closed-form solution θ_0^* to the equation $\frac{\partial}{\partial \theta_0} \log p(\mathbf{y} \mid X, \theta) = 0$:

○ $\frac{1}{m}\mathbf{y}^\top \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{y}$ ○ $\frac{1}{m} \operatorname{tr}(\tilde{\mathbf{K}}^{-1})$ √ $\frac{1}{m}\mathbf{y}^\top \tilde{\mathbf{K}}^{-1} \mathbf{y}$ ○ $\mathbf{y}^\top \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{y}$

Solution: We first define $\tilde{\mathbf{K}}$ the covariance matrix of $\mathbf{y}$ for covariance function $\tilde{k}$. Then, because $\mathbf{K}_{\mathbf{y},\theta} = \theta_0 \tilde{\mathbf{K}}$ and for any non-zero scalar s and invertible matrix A, we know $(s \times A)^{-1} = \frac{A^{-1}}{s}$, we have that:

$$\begin{aligned}\frac{\partial}{\partial \theta_0} \log p(\mathbf{y} \mid X, \boldsymbol{\theta}) &= \frac{1}{2} \operatorname{tr}\left(\mathbf{K}_{\mathbf{y},\theta}^{-1} \mathbf{y} (\mathbf{K}_{\mathbf{y},\theta}^{-1} \mathbf{y})^\top \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_0} - \mathbf{K}_{\mathbf{y},\theta}^{-1} \frac{\partial \mathbf{K}_{\mathbf{y},\theta}}{\partial \theta_0}\right) \\ &= \frac{1}{2} \operatorname{tr}\left(\left(\theta_0^{-2} \tilde{\mathbf{K}}^{-1} \mathbf{y} (\tilde{\mathbf{K}}^{-1} \mathbf{y})^\top - \theta_0^{-1} \tilde{\mathbf{K}}^{-1}\right) \tilde{\mathbf{K}}\right)\end{aligned}$$

Simplifying terms and using trace linearity, we obtain that :

$$\frac{\partial}{\partial \theta_0} \log p(\mathbf{y} \mid X, \boldsymbol{\theta}) = 0 \iff \theta_0^* = \frac{1}{m} \operatorname{tr}(\mathbf{y}\mathbf{y}^\top \tilde{\mathbf{K}}^{-1}) = \frac{1}{m}\mathbf{y}^\top \tilde{\mathbf{K}}^{-1} \mathbf{y}$$

If we define $\tilde{\mathbf{P}} = \tilde{\mathbf{K}}^{-1}$ as the precision matrix associated to $\mathbf{y}$ for covariance function $\tilde{k}$, we can rewrite θ_0^* in closed form:

$$\theta_0^* = \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{m} y_i \tilde{P}_{ik} y_k$$

(c) [☑]In the noise-free case, given the covariance function defined as $k_{\theta_0}(x, x') = \theta_0 \tilde{k}(x, x')$ with $\tilde{k}$ a valid kernel independent of θ_0 , how should we scale optimal parameter θ_0^* if we scale labels $\mathbf{y}$ by a scalar s:

✓ s^2
○ θ_0^* stays identical because it is independent of s
○ s
○ It depends of the choice of kernel function $\tilde{k}$

Solution: Following previous question, we find that θ_0^* depends quadratically on $\mathbf{y}$.
In particular:

$$\theta_0^* = \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{m} y_i \tilde{P}_{ik} y_k$$

Hence if $\tilde{\mathbf{y}} = s\mathbf{y}$ then:

$$\begin{aligned} \tilde{\theta_0^*} &= \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{m} \tilde{y}_i \tilde{P}_{ik} \tilde{y}_k \\ &= \frac{s^2}{m} \sum_{i=1}^{m} \sum_{k=1}^{m} y_i \tilde{P}_{ik} y_k \\ &= s^2 \theta_0^* \end{aligned}$$

Exercise 5: Equivalence between a subset of regressor and Gaussian process regression

Consider an unknown function $f : \mathcal{X} \to \mathbb{R}$. A priori we assume $f \sim GP(0, k)$. Using Gaussian process regression, after a set of $\mathbf{y}$ samples, the predictive mean $\mathbb{E}[f(x')]$ and variance $\mathbb{V}[f(x')]$ at x' are:

$$\mathbb{E}[f(x')] = \mathbf{k}(x')^\top \left(K + \sigma_n^2 I\right)^{-1} \mathbf{y} \tag{1}$$

$$\mathbb{V}[f(x')] = k(x', x') - \mathbf{k}(x')^\top (K + \sigma_n^2 I)^{-1} \mathbf{k}(x') \tag{2}$$

Unfortunately, the Gaussian process predictions scales typically $\mathcal{O}(n^3)$ due to matrix inversion. While using only $m < n$ subset of regressors helps to reduce the time complexity to $\mathcal{O}(m^2 n)$ with approximated predictive mean $\mathbb{E}[\tilde{f}(x')]$ and variance $\mathbb{V}[\tilde{f}(x')]$ at x' given by,

$$\mathbb{E}[\tilde{f}(x')] = \mathbf{k}_m(x')^\top \left(K_{mn} K_{nm} + \sigma_n^2 K_{mm}\right)^{-1} K_{mn} \mathbf{y} \tag{3}$$

$$\mathbb{V}[\tilde{f}(x')] = \sigma_n^2 \mathbf{k}_m(x')^\top \left(K_{mn} K_{nm} + \sigma_n^2 K_{mm}\right)^{-1} \mathbf{k}_m(x') \tag{4}$$

Show that subset of regressors predictors for the mean and variance is equivalent to the full Gaussian process regression predictors while using Nyström approximate kernel function $\tilde{k}(x, x') = \mathbf{k}_m^\top(x) K_{mm}^{-1} \mathbf{k}_m(x')$.

Hints: In order to show equivalence, one can write predictive mean $\mathbb{E}[\tilde{f}(x')]$ and variance $\mathbb{V}[\tilde{f}(x')]$ for the subset of regressors as predictors of Gaussian process regression (1, 2) but with approximated kernel function $\tilde{k}(x, x') = \mathbf{k}_m^\top(x) K_{mm}^{-1} \mathbf{k}_m(x')$ instead of $k(x, x')$.

Using the approximated kernel function $\tilde{k}(x, x')$for each pair of data points in the training set, we get, $\tilde{\mathbf{k}}(x') = K_{nm}K_{mm}^{-1}\mathbf{k}_m(x')$ and $\tilde{K} = K_{nm}K_{mm}^{-1}K_{mn}$.

For a $n \times m$ matrix, Q, you can use matrix inversion lemma, $(\sigma^2 I_n + QQ^\top)^{-1} = \sigma^{-2}I_n - \sigma^{-2}Q(\sigma^2 I_m + Q^\top Q)^{-1}Q^\top$, to transform the inversion of an $n \times n$ to inversion of a $m \times m$ matrix.

Solution: In order to show equivalence, one can write predictive mean $\mathbb{E}[\tilde{f}(x')]$ using 1 but with approximated kernel functions, $\tilde{\mathbf{k}}(x') = K_{nm}K_{mm}^{-1}\mathbf{k}_m(x')$ and $\tilde{K} = K_{nm}K_{mm}^{-1}K_{mn}$.

$$\begin{aligned}
\mathbb{E}[\tilde{f}(x')] &= \tilde{\mathbf{k}}(x')^\top \left(\tilde{K} + \sigma_n^2 I\right)^{-1} \mathbf{y} \\
&= \mathbf{k}_m^\top(x')K_{mm}^{-1}K_{mn}\left(K_{nm}K_{mm}^{-1}K_{mn} + \sigma_n^2 I\right)^{-1}\mathbf{y} \\
&= \sigma_n^{-2}\mathbf{k}_m^\top(x')K_{mm}^{-1}K_{mn}\left(I_n - K_{nm}\left(\sigma_n^2 K_{mm} + K_{mn}K_{nm}\right)^{-1}K_{mn}\right)\mathbf{y} \\
&= \sigma_n^{-2}\mathbf{k}_m^\top(x')K_{mm}^{-1}\left(I_m - K_{mn}K_{nm}\left(\sigma_n^2 K_{mm} + K_{mn}K_{nm}\right)^{-1}\right)K_{mn}\mathbf{y} \\
&= \sigma_n^{-2}\mathbf{k}_m^\top(x')K_{mm}^{-1}\left(\left(\sigma_n^2 K_{mm} + K_{mn}K_{nm}\right) - K_{mn}K_{nm}\right)\left(\sigma_n^2 K_{mm} + K_{mn}K_{nm}\right)^{-1}K_{mn}\mathbf{y} \\
&= \mathbf{k}_m^\top(x')\left(K_{mn}K_{nm} + \sigma_n^2 K_{mm}\right)^{-1}K_{mn}\mathbf{y}
\end{aligned} \tag{5}$$

Thus, we obtain Equation 3 which is a mean predictor for the subset of regressors. Equation 5 is obtained using matrix inversion lemma. Similarly, predictive variance $\mathbb{V}[\tilde{f}(x')]$ using 2 but with approximated kernel functions is given by,

$$\begin{aligned}
\mathbb{V}[\tilde{f}(x')] &= \tilde{k}(x', x') - \tilde{\mathbf{k}}(x')^\top(\tilde{K} + \sigma_n^2 I)^{-1}\tilde{\mathbf{k}}(x') \\
&= \mathbf{k}_m^\top(x')K_{mm}^{-1}\mathbf{k}_m(x') - \mathbf{k}_m^\top(x')K_{mm}^{-1}K_{mn}\left(K_{nm}K_{mm}^{-1}K_{mn} + \sigma_n^2 I\right)^{-1}K_{nm}K_{mm}^{-1}\mathbf{k}_m(x') \\
&= \mathbf{k}_m^\top(x')K_{mm}^{-1}\mathbf{k}_m(x') - \mathbf{k}_m(x')^\top\left(K_{mn}K_{nm} + \sigma_n^2 K_{mm}\right)^{-1}K_{mn}K_{nm}K_{mm}^{-1}\mathbf{k}_m(x') \\
&= \mathbf{k}_m^\top(x')\left(I_m - \left(K_{mn}K_{nm} + \sigma_n^2 K_{mm}\right)^{-1}K_{mn}K_{nm}\right)K_{mm}^{-1}\mathbf{k}_m(x') \\
&= \mathbf{k}_m^\top(x')K_{mm}^{-1}K_{mn}\left(K_{nm}K_{mm}^{-1}K_{mn} + \sigma_n^2 I\right)^{-1}\mathbf{k}_m(x') \\
&= \mathbf{k}_m^\top(x')\left(K_{mn}K_{nm} + \sigma_n^2 K_{mm}\right)^{-1}\left(\left(K_{mn}K_{nm} + \sigma_n^2 K_{mm}\right) - K_{mn}K_{nm}\right)K_{mm}^{-1}\mathbf{k}_m(x') \\
&= \sigma_n^2\mathbf{k}_m^\top(x')\left(K_{mn}K_{nm} + \sigma_n^2 K_{mm}\right)^{-1}K_{mm}K_{mm}^{-1}\mathbf{k}_m(x') \\
&= \sigma_n^2\mathbf{k}_m^\top(x')\left(K_{mn}K_{nm} + \sigma_n^2 K_{mm}\right)^{-1}\mathbf{k}_m(x')
\end{aligned} \tag{6}$$

Thus, we obtain Equation 4 which is a variance predictor for the subset of regressors. Equation 6 is obtained by following the same steps used to derive $\mathbb{E}[\tilde{f}(x')]$ in the first part.