Exercise sheet: Gaussian processes

The following exercises have different levels of difficulty indicated by (*), (**), (***). An exercise with (*) is a simple exercise requiring less time to solve compared to an exercise with (***), which is a more complex exercise.

1. (*) Let $f(t) = \int_0^t u(\tau)d\tau$. If $u(t) \sim \mathcal{GP}(0, k_u(t,t'))$, i.e. $u(t)$ is a GP with kernel function $k_u(t,t')$, write the expression that corresponds to the kernel function for $f(t)$, i.e. $k_f(t,t')$.
   Answer:
   The covariance function for $f(t)$ is defined as
$$k_f(t,t') = \mathrm{E}[f(t)f(t')] - \mathrm{E}[f(t)]\mathrm{E}[f(t)'],$$
   where $\mathrm{E}[f(t)] = \mathrm{E}[\int_0^t u(\tau)d\tau] = \int_0^t \mathrm{E}[u(\tau)]d\tau = 0$. Leading to
$$\begin{aligned}k_f(t,t') = \mathrm{E}[f(t)f(t')] = \mathrm{E}\left[\int_0^t u(\tau)d\tau \int_0^{t'} u(\tau')d\tau'\right] = \int_0^t\int_0^{t'} \mathrm{E}[u(\tau)u(\tau')]d\tau d\tau' \\ = \int_0^t\int_0^{t'} k_u(\tau,\tau')d\tau d\tau'.\end{aligned}$$

2. (*) The linear kernel is defined as $k(\mathbf{x},\mathbf{z}) = \mathbf{x}^\top\mathbf{z}$. If $\mathbf{X}$ is a design matrix of input vectors,
$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top\end{bmatrix},$$
   write the expression for the kernel matrix $\mathbf{K}$ in terms of the matrix $\mathbf{X}$.
   Answer:
   The kernel matrix $\mathbf{K}$ is defined as
$$\begin{aligned}\mathbf{K} = \begin{bmatrix}k(\mathbf{x}_1,\mathbf{x}_1) & k(\mathbf{x}_1,\mathbf{x}_2) & \cdots & k(\mathbf{x}_1,\mathbf{x}_N) \\ k(\mathbf{x}_2,\mathbf{x}_1) & k(\mathbf{x}_2,\mathbf{x}_2) & \cdots & k(\mathbf{x}_2,\mathbf{x}_N) \\ \vdots & \cdots & \cdots & \vdots \\ k(\mathbf{x}_N,\mathbf{x}_1) & k(\mathbf{x}_N,\mathbf{x}_2) & \cdots & k(\mathbf{x}_N,\mathbf{x}_N)\end{bmatrix} = \begin{bmatrix}\mathbf{x}_1^\top\mathbf{x}_1 & \mathbf{x}_1^\top\mathbf{x}_2 & \cdots & \mathbf{x}_1^\top\mathbf{x}_N \\ \mathbf{x}_2^\top\mathbf{x}_1 & \mathbf{x}_2^\top\mathbf{x}_2 & \cdots & \mathbf{x}_2^\top\mathbf{x}_N \\ \vdots & \cdots & \cdots & \vdots \\ \mathbf{x}_N^\top\mathbf{x}_1 & \mathbf{x}_N^\top\mathbf{x}_2 & \cdots & \mathbf{x}_N^\top\mathbf{x}_N\end{bmatrix} \\ = \begin{bmatrix}\mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N\end{bmatrix} = \mathbf{X}\mathbf{X}^\top\end{aligned}$$

3. (**) Using the properties for the marginal and conditional Gaussians (see Appendix A below) show that the posterior distribution for $p(\mathbf{w}|\mathbf{y},\mathbf{X})$ is given as
$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \mathcal{N}(\mathbf{w}|\frac{1}{\sigma_n^2}\mathbf{A}^{-1}\mathbf{\Phi}^\top\mathbf{y}, \mathbf{A}^{-1}),$$

where $\mathbf{A} = \sigma_n^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{\Sigma}_p^{-1}$, with $\mathbf{\Phi} \in \mathbb{R}^{n\times N}$.

Answer:

The likelihood is given as $p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{\Phi}\mathbf{w}, \sigma_n^2\mathbf{I})$ and the prior as $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{\Sigma}_p)$. We need to compute $p(\mathbf{w}|\mathbf{y},\mathbf{X})$. Looking at the equations for the marginal and conditional Gaussians, and assuming $\mathbf{w}$ replaces $\mathbf{x}$ in the appendix, i.e.

$$\boldsymbol{\mu} = \mathbf{0}, \quad \mathbf{\Lambda}^{-1} = \mathbf{\Sigma}_p, \quad \mathbf{B} = \mathbf{\Phi}, \quad \mathbf{b} = \mathbf{0}, \quad \mathbf{L}^{-1} = \sigma_n^2\mathbf{I},$$

we then have

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{\Sigma}\mathbf{\Phi}^\top\sigma_n^{-2}\mathbf{y}, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma} = (\mathbf{\Sigma}_p^{-1} + \mathbf{\Phi}^\top\sigma_n^{-2}\mathbf{\Phi})^{-1}$, which is the result we are looking for if we name $\mathbf{\Sigma}$ as $\mathbf{A}^{-1}$, leading to

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \mathcal{N}(\mathbf{w}|\frac{1}{\sigma_n^2}\mathbf{A}^{-1}\mathbf{\Phi}^\top\mathbf{y}, \mathbf{A}^{-1}),$$

where $\mathbf{A}^{-1} = (\sigma_n^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{\Sigma}_p^{-1})^{-1}$ or $\mathbf{A} = \sigma_n^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{\Sigma}_p^{-1}$.

4. (*) Show that the predictive distribution $p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y})$ is given as

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}\left(f_*\middle|\frac{1}{\sigma_n^2}\boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbf{A}^{-1}\mathbf{\Phi}^\top\mathbf{y}, \boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbf{A}^{-1}\boldsymbol{\phi}(\mathbf{x}_*)\right),$$

where $\mathbf{A} = \sigma_n^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{\Sigma}_p^{-1}$.

Answer:

To compute f_*, we use $f_*(\mathbf{x}_*) = \boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbf{w}$. The uncertainty in f_* comes from the uncertainty on $\mathbf{w}$ given by

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \mathcal{N}(\mathbf{w}|\frac{1}{\sigma_n^2}\mathbf{A}^{-1}\mathbf{\Phi}^\top\mathbf{y}, \mathbf{A}^{-1}),$$

where $\mathbf{A} = \sigma_n^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{\Sigma}_p^{-1}$. Since $\mathbf{w}$ is a Gaussian variable and $\boldsymbol{\phi}(\mathbf{x}_*)$ is a constant, $f_*(\mathbf{x}_*)$ is also a Gaussian, with mean and covariance given as

$$\begin{aligned}
\mathbb{E}[f_*(\mathbf{x}_*)] &= \mathbb{E}[\boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbf{w}] = \boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbb{E}[\mathbf{w}] = \boldsymbol{\phi}(\mathbf{x}_*)^\top\frac{1}{\sigma_n^2}\mathbf{A}^{-1}\mathbf{\Phi}^\top\mathbf{y} = \frac{1}{\sigma_n^2}\boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbf{A}^{-1}\mathbf{\Phi}^\top\mathbf{y}\\
\mathrm{var}[f_*(\mathbf{x}_*)]] &= \mathbb{E}[f_*(\mathbf{x}_*)f_*^\top(\mathbf{x}_*)] - \mathbb{E}[f_*(\mathbf{x}_*)]\mathbb{E}[f_*^\top(\mathbf{x}_*)] = \mathbb{E}[\boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbf{w}\mathbf{w}^\top\boldsymbol{\phi}(\mathbf{x}_*)] - \mathbb{E}[\boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbf{w}]\mathbb{E}[\mathbf{w}^\top\boldsymbol{\phi}(\mathbf{x}_*)]\\
&= \boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbb{E}[\mathbf{w}\mathbf{w}^\top]\boldsymbol{\phi}(\mathbf{x}_*) - \boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbb{E}[\mathbf{w}]\mathbb{E}[\mathbf{w}^\top]\boldsymbol{\phi}(\mathbf{x}_*) = \boldsymbol{\phi}(\mathbf{x}_*)^\top\left[\mathbb{E}[\mathbf{w}\mathbf{w}^\top] - \mathbb{E}[\mathbf{w}]\mathbb{E}[\mathbf{w}^\top]\right]\boldsymbol{\phi}(\mathbf{x}_*)\\
&= \boldsymbol{\phi}(\mathbf{x}_*)^\top\,\mathrm{cov}[\mathbf{w}]\boldsymbol{\phi}(\mathbf{x}_*) = \boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbf{A}^{-1}\boldsymbol{\phi}(\mathbf{x}_*).
\end{aligned}$$

Therefore,

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(f_*|\frac{1}{\sigma_n^2}\boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbf{A}^{-1}\mathbf{\Phi}^\top\mathbf{y}, \boldsymbol{\phi}(\mathbf{x}_*)^\top\mathbf{A}^{-1}\boldsymbol{\phi}(\mathbf{x}_*)).$$

5. (**) Show that another way to write the predictive distribution from the previous exercise is

$$\begin{aligned}
p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}\Big(f_*\Big|&\boldsymbol{\phi}_*^\top\mathbf{\Sigma}_p\mathbf{\Phi}^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y},\\
&\boldsymbol{\phi}_*^\top\mathbf{\Sigma}_p\boldsymbol{\phi}_* - \boldsymbol{\phi}_*^\top\mathbf{\Sigma}_p\mathbf{\Phi}^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{\Phi}\mathbf{\Sigma}_p\boldsymbol{\phi}_*\Big),
\end{aligned}$$

where $\boldsymbol{\phi}(\mathbf{x}_*) = \boldsymbol{\phi}_*$, y $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top$.
[HINT: use the properties for the matrix inverses shown in Appendix B]
Answer:
Let us start with the mean of the predictive distribution $\frac{1}{\sigma_n^2}\boldsymbol{\phi}(\mathbf{x}_*)^\top \mathbf{A}^{-1}\boldsymbol{\Phi}^\top \mathbf{y}$,

$$\begin{aligned}\frac{1}{\sigma_n^2}\boldsymbol{\phi}(\mathbf{x}_*)^\top \mathbf{A}^{-1}\boldsymbol{\Phi}^\top \mathbf{y} &= \boldsymbol{\phi}(\mathbf{x}_*)^\top \left[\sigma_n^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{\Sigma}_p^{-1}\right]^{-1} \boldsymbol{\Phi}^\top \sigma_n^{-2}\mathbf{y} \\ &= \boldsymbol{\phi}(\mathbf{x}_*)^\top \underbrace{\left[\boldsymbol{\Sigma}_p^{-1} + \boldsymbol{\Phi}^\top\sigma_n^{-2}\mathbf{I}\boldsymbol{\Phi}\right]^{-1} \boldsymbol{\Phi}^\top \sigma_n^{-2}\mathbf{I}}_{U}\,\mathbf{y},\end{aligned}$$

where we have re-organised some terms. For the expression in U above, we can use the first identity matrix in Appendix B, assuming

$$\mathbf{P} = \boldsymbol{\Sigma}_p, \quad \mathbf{B} = \boldsymbol{\Phi}, \quad \mathbf{R} = \sigma_n^2\mathbf{I}.$$

This leads to

$$\left[\boldsymbol{\Sigma}_p^{-1} + \boldsymbol{\Phi}^\top\sigma_n^{-2}\mathbf{I}\boldsymbol{\Phi}\right]^{-1} \boldsymbol{\Phi}^\top \sigma_n^{-2}\mathbf{I} = \boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top \left[\boldsymbol{\Phi}\boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top + \sigma_n^2\mathbf{I}\right]^{-1}.$$

Leading to the following mean for the updated predictive distribution,

$$\boldsymbol{\phi}(\mathbf{x}_*)^\top \boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top \left[\boldsymbol{\Phi}\boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top + \sigma_n^2\mathbf{I}\right]^{-1}\mathbf{y} = \boldsymbol{\phi}(\mathbf{x}_*)^\top \boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top \left[\mathbf{K} + \sigma_n^2\mathbf{I}\right]^{-1}\mathbf{y},$$

where $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top$.

For the case of the predictive variance $\boldsymbol{\phi}(\mathbf{x}_*)^\top \mathbf{A}^{-1}\boldsymbol{\phi}(\mathbf{x}_*)$, we can write

$$\boldsymbol{\phi}(\mathbf{x}_*)^\top \mathbf{A}^{-1}\boldsymbol{\phi}(\mathbf{x}_*) = \boldsymbol{\phi}(\mathbf{x}_*)^\top \underbrace{\left[\boldsymbol{\Sigma}_p^{-1} + \boldsymbol{\Phi}^\top\sigma_n^{-2}\mathbf{I}\boldsymbol{\Phi}\right]^{-1}}_{U}\boldsymbol{\phi}(\mathbf{x}_*).$$

For the U term above, we can apply the Woodbury identity of Appendix B assuming

$$\mathbf{A} = \boldsymbol{\Sigma}_p^{-1}, \quad \mathbf{B} = \boldsymbol{\Phi}^\top, \quad \mathbf{D} = \sigma_n^2\mathbf{I}, \quad \mathbf{C} = \boldsymbol{\Phi}.$$

This leads to

$$\begin{aligned}\left[\boldsymbol{\Sigma}_p^{-1} + \boldsymbol{\Phi}^\top\sigma_n^{-2}\mathbf{I}\boldsymbol{\Phi}\right]^{-1} &= \boldsymbol{\Sigma}_p - \boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top(\sigma_n^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top)^{-1}\boldsymbol{\Phi}\boldsymbol{\Sigma}_p \\ &= \boldsymbol{\Sigma}_p - \boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top + \sigma_n^2\mathbf{I})^{-1}\boldsymbol{\Phi}\boldsymbol{\Sigma}_p \\ &= \boldsymbol{\Sigma}_p - \boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\boldsymbol{\Phi}\boldsymbol{\Sigma}_p.\end{aligned}$$

The predictive variance is then

$$\begin{aligned}\boldsymbol{\phi}(\mathbf{x}_*)^\top \mathbf{A}^{-1}\boldsymbol{\phi}(\mathbf{x}_*) &= \boldsymbol{\phi}(\mathbf{x}_*)^\top \left[\boldsymbol{\Sigma}_p - \boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\boldsymbol{\Phi}\boldsymbol{\Sigma}_p\right]\boldsymbol{\phi}(\mathbf{x}_*) \\ &= \boldsymbol{\phi}(\mathbf{x}_*)^\top\boldsymbol{\Sigma}_p\boldsymbol{\phi}(\mathbf{x}_*) - \boldsymbol{\phi}(\mathbf{x}_*)^\top\boldsymbol{\Sigma}_p\boldsymbol{\Phi}^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\boldsymbol{\Phi}\boldsymbol{\Sigma}_p\boldsymbol{\phi}(\mathbf{x}_*).\end{aligned}$$

6. (*) Show that if $k_1(\mathbf{x}, \mathbf{x}')$ is a valid kernel, then $k(\mathbf{x}, \mathbf{x}') = ck_1(\mathbf{x}, \mathbf{x}')$, with $c > 0$ is a valid kernel.
Answer:
We saw in the Lecture that the kernel trick defines a valid kernel as the inner product between two vectors,

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\psi}(\mathbf{x}).$$

It follows that

$$ck_1\left(\mathbf{x}, \mathbf{x}'\right) = \mathbf{u}(\mathbf{x})^\top \mathbf{u}\left(\mathbf{x}'\right),$$

where $\mathbf{u}(\mathbf{x}) = c^{1/2}\boldsymbol{\psi}(\mathbf{x})$, and so $ck_1\left(\mathbf{x}, \mathbf{x}'\right)$ can be expressed as the inner product of two feature vectors, and therefore is a valid kernel.

7. (*) Show that $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{A}\mathbf{x}'$ is a valid kernel, with $\mathbf{A}$ a symmetric positive semidefinite matrix.
Answer:
Let the kernel matrix be $\mathbf{K} = \mathbf{X}^\top \mathbf{A}\mathbf{X}$, where the kernel matrix has entries $(\mathbf{K})_{i,j} = \mathbf{x}_i^\top \mathbf{A}\mathbf{x}_j$. Consider

$$\begin{aligned} \mathbf{u}^\top \mathbf{K}\mathbf{u} &= \mathbf{u}^\top \mathbf{X}^\top \mathbf{A}\mathbf{X}\mathbf{u} \\ &= \mathbf{v}^\top \mathbf{A}\mathbf{v} \geqslant 0, \end{aligned}$$

where $\mathbf{v} = \mathbf{X}\mathbf{u}$, and we have use the fact that $\mathbf{A}$ is a symmetric positive semidefinite matrix.

8. (**) Let $\text{var}_n(f(\mathbf{x}_*))$ be the predictive variance of a Gaussian process regression model at $\mathbf{x}_*$ given a dataset of size n. The corresponding predictive variance using a dataset of only the first $n-1$ training points is denoted $\text{var}_{n-1}(f(\mathbf{x}_*))$. Show that $\text{var}_n(f(\mathbf{x}_*)) \leq \text{var}_{n-1}(f(\mathbf{x}_*))$, i.e. that the predictive variance at $\mathbf{x}_*$ cannot increase as more training data is obtained.
[HINT: use the inverse of a partitioned matrix as shown in Appendix B]
Answer:
The predictive variance $\text{var}_n(f(\mathbf{x}_*))$ is given as

$$\text{var}_n(f(\mathbf{x}_*)) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n})\mathbf{K}_n^{-1}\mathbf{k}(\mathbf{X}_{1:n}, \mathbf{x}_*),$$

where

$$\mathbf{K}_n = \begin{bmatrix} \mathbf{K}_{n-1} & \mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n) \\ \mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1}) & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}.$$

Using the inverse of a partitioned matrix in Appendix B,

$$\mathbf{A} = \begin{pmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{R} & \mathbf{S} \end{pmatrix}, \quad \mathbf{A}^{-1} = \begin{pmatrix} \tilde{\mathbf{P}} & \tilde{\mathbf{Q}} \\ \tilde{\mathbf{R}} & \tilde{\mathbf{S}} \end{pmatrix},$$

where

$$\left.\begin{aligned} \tilde{\mathbf{P}} &= \mathbf{P}^{-1} + \mathbf{P}^{-1}\mathbf{Q}\mathbf{M}\mathbf{R}\mathbf{P}^{-1} \\ \tilde{\mathbf{Q}} &= -\mathbf{P}^{-1}\mathbf{Q}\mathbf{M} \\ \tilde{\mathbf{R}} &= -\mathbf{M}\mathbf{R}\mathbf{P}^{-1} \\ \tilde{\mathbf{S}} &= \mathbf{M} \end{aligned}\right\} \text{ where } \mathbf{M} = \left(\mathbf{S} - \mathbf{R}\mathbf{P}^{-1}\mathbf{Q}\right)^{-1}$$

and assuming

$$\mathbf{P} = \mathbf{K}_{n-1}, \quad \mathbf{Q} = \mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n), \quad \mathbf{R} = \mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1}), \quad \mathbf{S} = k(\mathbf{x}_n, \mathbf{x}_n),$$

we can compute the inverse for $\mathbf{K}_n$ as

$$\begin{aligned}
\mathbf{K}_n^{-1} &= \begin{bmatrix} \mathbf{K}_{n-1} & \mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n) \\ \mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1}) & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \mathbf{K}_{n-1}^{-1} + \mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1} & -\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M} \\ -\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1} & \mathbf{M} \end{bmatrix},
\end{aligned}$$

where $\mathbf{M} = \big(k(\mathbf{x}_n, \mathbf{x}_n) - \mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\big)^{-1}$.

The predictive variance $\mathrm{var}_n(f(\mathbf{x}_*))$ follows as

$$\begin{aligned}
\mathrm{var}_n(f(\mathbf{x}_*)) &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n})\mathbf{K}_n^{-1}\mathbf{k}(\mathbf{X}_{1:n}, \mathbf{x}_*) \\
&= k(\mathbf{x}_*, \mathbf{x}_*) - \Big[\mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1}), k(\mathbf{x}_*, \mathbf{x}_n)\Big] * \\
&\begin{bmatrix} \mathbf{K}_{n-1}^{-1} + \mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1} & -\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M} \\ -\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1} & \mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{x}_n) \end{bmatrix}.
\end{aligned}$$

The second term in the rhs in the expression above follows as

$$\begin{bmatrix} \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1} + \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1} - k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1} \\ -\mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M} + k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M} \end{bmatrix}^\top \begin{bmatrix} \mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{x}_n) \end{bmatrix}$$

following as

$$\begin{aligned}
&\mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&+ \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&- k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&- \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}k(\mathbf{x}_*, \mathbf{x}_n) \\
&+ k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}k(\mathbf{x}_*, \mathbf{x}_n)
\end{aligned}$$

The predictive variance $\mathrm{var}_n(f(\mathbf{x}_*))$ follows as

$$\begin{aligned}
\mathrm{var}_n(f(\mathbf{x}_*)) &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&- \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&+ k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&+ \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}k(\mathbf{x}_*, \mathbf{x}_n) \\
&- k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}k(\mathbf{x}_*, \mathbf{x}_n).
\end{aligned}$$

The predictive variance $\mathrm{var}_{n-1}(f(\mathbf{x}_*))$ is given as

$$\mathrm{var}_{n-1}(f(\mathbf{x}_*)) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*)$$

We need to show that

$$
\begin{aligned}
&\mathrm{var}_n(f(\mathbf{x}_*)) \leq \mathrm{var}_{n-1}(f(\mathbf{x}_*)) \\
&k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&- \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&+ k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&+ \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}k(\mathbf{x}_*, \mathbf{x}_n) \\
&- k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}k(\mathbf{x}_*, \mathbf{x}_n) \\
&\leq \\
&k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*)
\end{aligned}
$$

Or that

$$
\begin{aligned}
&- \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&+ k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&+ \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}k(\mathbf{x}_*, \mathbf{x}_n) \\
&- k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}k(\mathbf{x}_*, \mathbf{x}_n) \\
&\leq 0
\end{aligned}
$$

Or that

$$
\begin{aligned}
&k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) + \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}k(\mathbf{x}_*, \mathbf{x}_n) \\
&\leq \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) + k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}k(\mathbf{x}_*, \mathbf{x}_n)
\end{aligned}
$$

Or that

$$
\begin{aligned}
&2k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&\leq \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{M}\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) + k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{M}k(\mathbf{x}_*, \mathbf{x}_n)
\end{aligned}
$$

Since $\mathbf{M}$ is a scalar, we can divide both sides by $\mathbf{M}$ and we get

$$
\begin{aligned}
&2k(\mathbf{x}_*, \mathbf{x}_n)\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) \\
&\leq \mathbf{k}^\top(\mathbf{x}_*, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_n)\mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*) + k^2(\mathbf{x}_*, \mathbf{x}_n).
\end{aligned}
$$

Let us call $a = \mathbf{k}^\top(\mathbf{x}_n, \mathbf{X}_{1:n-1})\mathbf{K}_{n-1}^{-1}\mathbf{k}(\mathbf{X}_{1:n-1}, \mathbf{x}_*)$ and $b = k(\mathbf{x}_*, \mathbf{x}_n)$, meaning that

$$
\begin{aligned}
2ab &\leq a^2 + b^2 \\
0 &\leq a^2 - 2ab + b^2 \\
0 &\leq (a-b)^2,
\end{aligned}
$$

which follows for any value of a and b.

Appendix A: marginal and conditional Gaussians

Given a marginal Gaussian distribution for $\mathbf{x}$, and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$,

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{B}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}),$$

the marginal distribution for $\mathbf{y}$, and the conditional distribution for $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{B}\boldsymbol{\Lambda}^{-1}\mathbf{B}^\top)$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{B}^\top\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{B}^\top\mathbf{L}\mathbf{B})^{-1}.$$

Appendix B: matrix identities involving inverses

A useful identity involving matrix inverses is the following

$$\left(\mathbf{P}^{-1} + \mathbf{B}^\top\mathbf{R}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^\top\mathbf{R}^{-1} = \mathbf{P}\mathbf{B}^\top\left(\mathbf{B}\mathbf{P}\mathbf{B}^\top + \mathbf{R}\right)^{-1}.$$

Say $\mathbf{P} \in \mathbb{R}^{N\times N}$ and $\mathbf{R} \in \mathbb{R}^{M\times M}$, so that $\mathbf{B} \in \mathbb{R}^{M\times N}$. If $M \ll N$, it is much cheaper to evaluate the right-hand side of the expression above than the left-hand side.

Another useful identity involving inverses is the following:

$$\left(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1},$$

which is known as the *Woodbury identity*. This is useful, for instance, when $\mathbf{A}$ is large and diagonal, and hence easy to invert, while $\mathbf{B}$ has many rows but few columns (and conversely for $\mathbf{C}$) so that the right-hand side is much cheaper to evaluate than the left-hand side.

One more useful identity involving inverses is the following. Let the invertible $n \times n$ matrix $\mathbf{A}$ and its inverse $\mathbf{A}^{-1}$ be partitioned into

$$\mathbf{A} = \begin{pmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{R} & \mathbf{S} \end{pmatrix}, \quad \mathbf{A}^{-1} = \begin{pmatrix} \tilde{\mathbf{P}} & \tilde{\mathbf{Q}} \\ \tilde{\mathbf{R}} & \tilde{\mathbf{S}} \end{pmatrix},$$

where $\mathbf{P}$ and $\tilde{\mathbf{P}}$ are $n_1 \times n_1$ matrices and $\mathbf{S}$ and $\tilde{\mathbf{S}}$ are $n_2 \times n_2$ matrices with $n = n_1 + n_2$. The submatrices of $\mathbf{A}^{-1}$ are given

$$\left.\begin{aligned} \tilde{\mathbf{P}} &= \mathbf{P}^{-1} + \mathbf{P}^{-1}\mathbf{Q}\mathbf{M}\mathbf{R}\mathbf{P}^{-1} \\ \tilde{\mathbf{Q}} &= -\mathbf{P}^{-1}\mathbf{Q}\mathbf{M} \\ \tilde{\mathbf{R}} &= -\mathbf{M}\mathbf{R}\mathbf{P}^{-1} \\ \tilde{\mathbf{S}} &= \mathbf{M} \end{aligned}\right\} \text{ where } \mathbf{M} = \left(\mathbf{S} - \mathbf{R}\mathbf{P}^{-1}\mathbf{Q}\right)^{-1}$$