02477 Bayesian Machine Learning 2025: Assignment 2

This is the second assignment out of three in the Bayesian machine learning course 2025. The assignment is a group work of 3-5 students (please use the same groups as in assignment 1 if possible) and hand in via DTU Learn). The assignment is **mandatory**. The deadline is **6th of April 23:59**. The solution must be handed in as a single **PDF** document.

Part 1: Gaussian processes and covariance functions

In this part, we will study covariance functions for Gaussian process models. Consider the following six covariance functions

$$k_1(x, x') = 2\exp\left(-\frac{(x-x')^2}{2\cdot 0.3^2}\right) \tag{1}$$

$$k_2(x, x') = \exp\left(-\frac{(x-x')^2}{2\cdot 0.1^2}\right) \tag{2}$$

$$k_3(x, x') = 4 + 2xx' \tag{3}$$

$$k_4(x, x') = \exp\left(-2\sin\left(3\pi\cdot|x-x'|\right)^2\right) \tag{4}$$

$$k_5(x, x') = \exp\left(-2\sin\left(3\pi\cdot|x-x'|\right)^2\right) + 4xx' \tag{5}$$

$$k_6(x, x') = \frac{1}{5} + \min(x, x') \tag{6}$$

Some of them should be familiar and some of them might be new to you.

Task 1.1: Detertime the analytical marginal prior mean and variance of a Gaussian process, $f_i(x) \sim \mathcal{GP}(0, k_i(x, x'))$, i.e. compute $\mathbb{E}[f_i(x)]$ and $\mathbb{V}[f_i(x)]$ for each of the six covariance functions (for $i = 1, 2, \dots, 6$).

Task 1.2: Which of the six covariance functions are stationary covariance functions?

Task 1.3: Let $\boldsymbol{X} = \{x_i\}_{i=1}^{100}$ be a sorted set of equidistant points in the interval $[0, 2]$. Figure 1 shows the covariance matrices for function values evaluated at $\boldsymbol{X}$ as well as samples from the corresponding Gaussian process prior for each of the six covariance functions. Match the plots to each of the six covariance functions.

Consider now the following kernel function:

$$k_7(x, x') = \kappa_0^2 + \kappa_1^2 xx' + \kappa_2^2\exp\left(-\frac{||x-x'||_2^2}{2\ell^2}\right), \tag{7}$$

where $\kappa_1, \kappa_2 \geq 0$ and $\ell > 0$ are hyperparameters.

Task 1.4: Implement the kernel function k_7. Generate and plot $S = 30$ realizations (samples) of the process $f(x) \sim \mathcal{GP}(0, k_7(x, x'))$ for $x \in [-6, 6]$ for $(\kappa_0, \kappa_1, \kappa_2, \ell) = (5, 2, 0, \frac{1}{2})$. Repeat with $(\kappa_0, \kappa_1, \kappa_2, \ell) = (5, 0, 1, \frac{1}{2})$ and $(\kappa_0, \kappa_1, \kappa_2, \ell) = (5, 2, 1, \frac{1}{2})$

Hint: You can either adapt the code template from the GP exercises or you can simply implement the relevant equations directly.

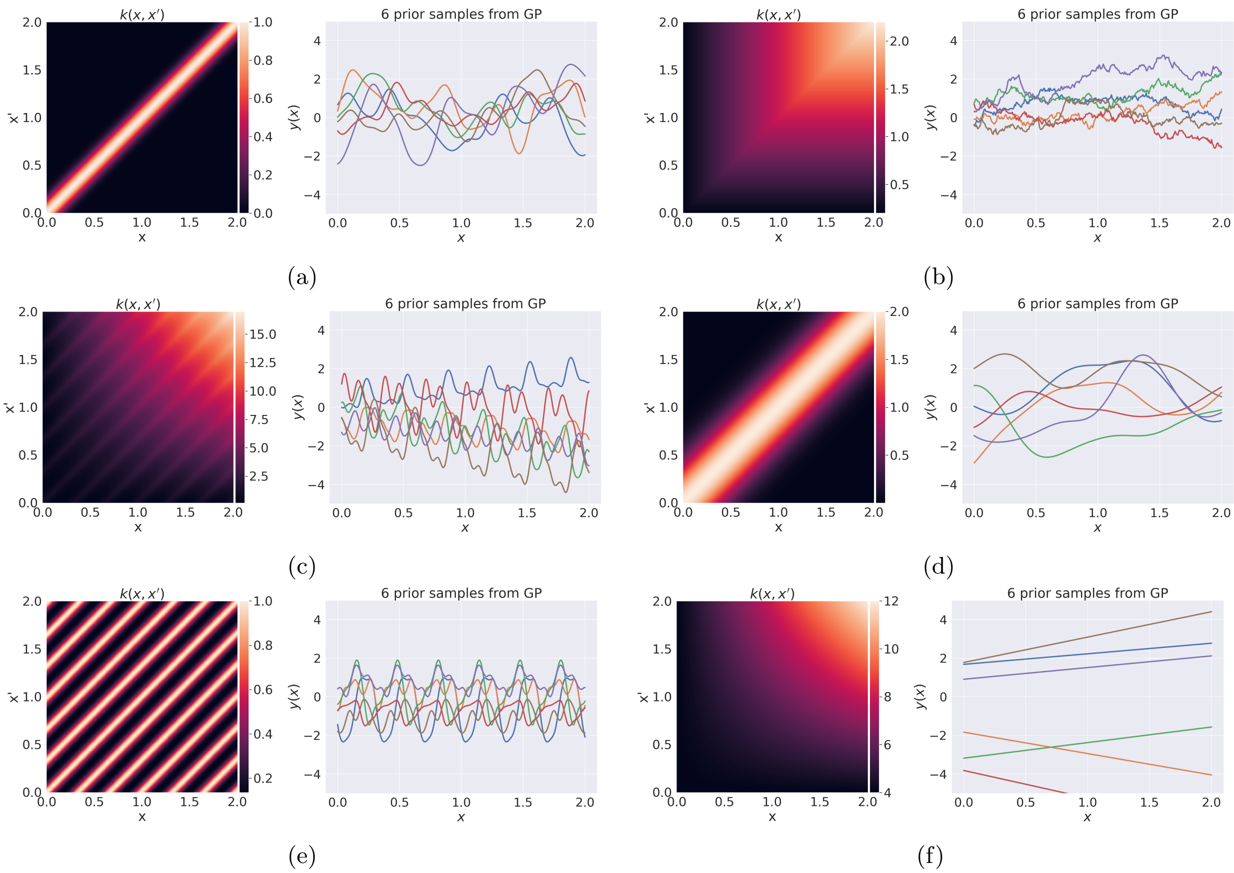

Figure 1: Covariance matrices and samples from the corresponding Gaussian process prior distribution for the six different covariance functions in randomized order.

Part 2: Laplace approximation for a simple neural network

Consider the following small regression data set with $N = 10$ observations $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, where

$$\boldsymbol{x} = \begin{bmatrix} 9.589, 7.375, 4.647, 2.501, 2.538, 6.783, 4.294, 5.111, 0.130, & 0.783 \end{bmatrix} \tag{8}$$

$$\boldsymbol{y} = \begin{bmatrix} 3.032, 3.349, 2.906, 2.126, 1.538, 2.787, 3.078, 2.993, 0.828, -0.331 \end{bmatrix} \tag{9}$$

such that the n'th element in $\boldsymbol{x}$ is $x_n \in \mathbb{R}$ and similar for $y_n \in \mathbb{R}$.

We will work with the following very simple two-parameter neural network:

$$f(x) = w_1 \sigma(x + w_0), \tag{10}$$

where $\boldsymbol{w} = \begin{bmatrix} w_0 & w_1 \end{bmatrix} \in \mathbb{R}^2$ is the parameters of the network and $\sigma : \mathbb{R} \rightarrow (0, 1)$ is the logistic sigmoid function. We can construct a non-linear probabilistic model for regression by using the network $f(x)$ as the mean function of a Gaussian likelihood:

$$p(y_n | x_n, \boldsymbol{w}) = \mathcal{N}(y_n | f(x), \beta^{-1}) = \mathcal{N}(y_n | w_1 \sigma(x + w_0), \beta^{-1}), \tag{11}$$

where $\beta^{-1} > 0$ is the noise variance. We impose i.i.d. Gaussian prior distributions on both parameters:

$$w_1 \sim \mathcal{N}(0, \tau^2) \tag{12}$$

$$w_2 \sim \mathcal{N}(0, \tau^2), \tag{13}$$

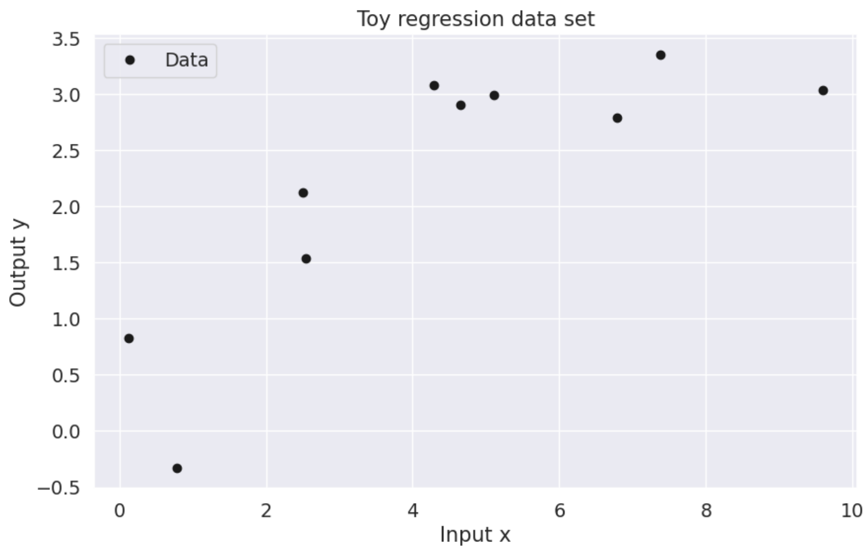

Figure 2: Data for Part 2

where $\tau^2 > 0$ is variance of the prior.

Assume $\tau = 2$ and $\beta = 4$.

Task 2.1: Generate $S = 100$ samples from the prior $p(\boldsymbol{w})$ and plot the corresponding functions $f(x)$ for $x \in [0, 10]$ on top of a scatter plot of the data.

Hint: You can find inspiration in the course notebooks in order to make the plots look nice.

The joint density for $(\boldsymbol{y}, \boldsymbol{w})$ according to the probabilistic model in eq. (11)-(13) is given

$$p(\boldsymbol{y}, \boldsymbol{w}) = p(\boldsymbol{w}) \prod_{n=1}^{N} p(y_n | x_n, \boldsymbol{w}) = \mathcal{N}(w_0 | 0, \tau^2) \mathcal{N}(w_1 | 0, \tau^2) \prod_{n=1}^{N} \mathcal{N}(y_n | w_1 \sigma(x_n + w_0), \beta^{-1}) \tag{14}$$

Task 2.2: What prevents us from using the equations in section 3.3 in Murphy1 to compute the exact posterior distribution of the parameters given the data analytically for the system in eq. (14) ?

Task 2.3: Implement a python for function for evaluating the *logarithm of the joint density* in eq. (14). Report the numerical value of $\log p(\boldsymbol{y}, \boldsymbol{w})$ for the dataset given above when $w_0 = w_1 = 0$.

Task 2.4: Determine the expression for the gradient and Hessian of the logarithm of the joint density wrt. $\boldsymbol{w}$.

Hints: 1) The chain rule will be handy. 2) Recall the identity $\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$.

Task 2.5: Determine the Laplace approximation $\boldsymbol{q}(\boldsymbol{w})$ of the posterior distribution such that $p(\boldsymbol{w}|\boldsymbol{y}) \approx \boldsymbol{q}(\boldsymbol{w})$.

Task 2.6: Create a 2D contour plot of posterior distribution of the parameters. Plot the contours of the Laplace approximation on top of the plot to validate your results from the previous tasks.

Hints: Find inspiration in the exercise notebook for week 3 & 4 if you are struggling with the plotting code.

Task 2.7: Generate and plot $S = 100$ samples from the approximate posterior distribution. Pot the corresponding function $f(x)$ for $x \in [0, 10]$ on top of a scatter plot of the data.

Task 2.8: Use the Laplace approximation to compute the (approximate) posterior probability for the event $f(8) > 3$.