1. Consider modelling the mean function \( m \) of the Gaussian process prior \( f \sim GP(m, k_\theta) \) with another GP: \( m \sim GP(0, k_\eta) \).

   (a) Show that this is equivalent to a zero-mean GP prior on \( f \) and find its covariance function.

   (b) Consider constraining the mean functions such that they follow a particular type of functions:

      (i) constant \( m(x) \equiv b \), with \( b \sim N(0, \sigma_b^2) \) (ii) linear \( m(x) = w^T x + b \), with \( w \sim N(0, \sigma_w^2 I) \) and \( b \sim N(0, \sigma_b^2) \) independent. Find the appropriate covariance functions \( k_\eta \).

2. Consider a GP regression model with \( f \sim GP(0, k) \) and \( y_i \sim N(f(x_i), \sigma^2) \). For training inputs \( x = \{x_i\}_{i=1}^n \) and outputs \( y = [y_1, \ldots, y_n]^T \) we denote the vector of evaluations of \( f \) by \( f = [f(x_1), \ldots, f(x_n)]^T \in \mathbb{R}^n \). We also have test inputs \( x_* = \{x_{*j}\}_{j=1}^m \) and denote the corresponding evaluations of \( f \) by \( f_* = [f(x_{*1}), \ldots, f(x_{*m})]^T \in \mathbb{R}^m \).

   (a) Write down the joint distribution of \( \begin{bmatrix} f \\ y \end{bmatrix} \) and thus compute \( p(f|y), p(f_*|f) \) and \( p(f_*|y). \)

   (b) Verify that \( p(f_*|y) = \int p(f_*|f)p(f|y)df \).

   [Hint: \( \int N(a|Bc, D)N(c|e, F)dc = N(a|Be, D + BF B^T) \)]

3. Consider a GP regression model in which the response variable \( y \) is \( d \)-dimensional, i.e. \( y \in \mathbb{R}^d \). Assuming that the individual response dimensions \( y^{(1)}, \ldots, y^{(d)} \) are conditionally independent given the input vector \( x \) with

\[
y^{(j)}|x \sim N(f^{(j)}(x), \lambda),
\]

with independent priors \( f^{(j)} \sim GP(0, k_\theta) \). Derive the posterior predictive distribution

\[
p(y_*|x_*, \{x_i, y_i\}_{i=1}^n),
\]

for a test input vector \( x_* \) and the training set \( \{x_i, y_i\}_{i=1}^n \).

Comment on the difference between this model and \( d \) independent Gaussian process regressions.

4. We observe \( \{(x_i, y_i)\}_{i=1}^n \), with \( x_i \in \mathbb{R}^p \) and \( y_i \in \{0, 1, 2, \ldots\} \). Consider a Gaussian process model with a Poisson link. Denoting \( f = [f(x_1), \ldots, f(x_n)] \), we have a prior \( f \sim N(0, K) \) and the likelihood

\[
p(y_i = r|f(x_i)) = \frac{e^{rf(x_i)} \exp(-e^{f(x_i)})}{r!}, \quad i = 1, \ldots, n,
\]

i.e. given \( f(x_i) \), \( y_i \) follows a Poisson distribution with rate \( \lambda(x_i) = e^{f(x_i)} \). We will assume that \( K \) is invertible.

   (a) Compute the log-posterior \( \log p(f|y) \) up to an additive constant and its gradient.

   (b) Compute the Hessian and verify that it is negative definite. Briefly describe how you would find a posterior mode \( \hat{f}_{\text{MAP}} \) of \( f \).

   (c) Construct a Laplace approximation to the posterior \( p(f|y) \) and compute the resulting approximation to the posterior predictive \( p(f(x_*)|y) \) for a new input \( x_* \). Compare it to the prediction \( p(f(x_*)|\hat{f}_{\text{MAP}}) \), based on the point estimate \( \hat{f}_{\text{MAP}} \) of \( f \). [Hint: you may find the following version of Woodbury identity useful: \( (A^{-1} + D)^{-1} = A - A(A + D^{-1})^{-1}A \) for invertible matrices \( A \) and \( D \)]
5. Suppose you have some frequencies $\omega_1, \ldots, \omega_m \sim \lambda$ to approximate a translation invariant kernel

$$k(x, x') = \kappa(\frac{x-x'}{\gamma}) = \int \exp \left( i \omega^\top (x - x') \right) \lambda(\omega) d\omega$$

with random Fourier features

$$\varphi_\omega(x) = \frac{1}{\sqrt{m}} \left[ \exp(i\omega_1^\top x), \ldots, \exp(i\omega_m^\top x) \right]$$

Assume you wish to double the lengthscale parameter $\gamma$. How would you modify the feature representation?

You also have frequencies $\eta_1, \ldots, \eta_m \sim \nu$ for another kernel $l(x, x') = \int \exp \left( i \eta^\top (x - x') \right) \nu(\eta) d\eta$.

Describe two ways to construct a feature map approximation of the product kernel $k(x, x')l(x, x')$.

6. (Ex. 24) In lecture notes on Bayesian optimization, we derived the probability of improvement and expected improvement acquisition function which ignore the noise in $\tilde{y}$. Derive the corrected versions.