SC4/SM8 Advanced Topics in Statistical Machine Learning
Problem Sheet 4

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

   (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 $\mathbf{m}(x) \equiv b$, with $b \sim \mathcal{N}(0, \sigma_b^2)$ (ii) linear $\mathbf{m}(x) = w^\top x + b$, with $w \sim \mathcal{N}(0, \sigma_w^2 I)$ and $b \sim \mathcal{N}(0, \sigma_b^2)$ independent. Find the appropriate covariance functions k_η.

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

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

   (b) Verify that $p(\mathbf{f}_\star|\mathbf{y}) = \int p(\mathbf{f}_\star|\mathbf{f}) p(\mathbf{f}|\mathbf{y}) d\mathbf{f}$.
   [*Hint*: $\int \mathcal{N}(a|Bc, D)\mathcal{N}(c|e, F) dc = \mathcal{N}(a|Be, D + BFB^\top)$]

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
$$y^{(j)}|x \sim \mathcal{N}(f^{(j)}(x), \lambda),$$
   with independent priors $f^{(j)} \sim \mathcal{GP}(0, k_\theta)$. Derive the posterior predictive distribution
$$p\left(y_\star | x_\star, \{x_i, y_i\}_{i=1}^n\right),$$
   for a test input vector $x_\star$ 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 $\mathbf{f} = [f(x_1), \ldots, f(x_n)]$, we have a prior $\mathbf{f} \sim \mathcal{N}(0, \mathbf{K})$ and the likelihood
$$p(y_i = r | f(x_i)) = \frac{e^{r f(x_i)} \exp(-e^{f(x_i)})}{r!}, \quad i = 1, \ldots, n, \tag{1}$$
   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 $\mathbf{K}$ is invertible.

   (a) Compute the log-posterior $\log p(\mathbf{f}|\mathbf{y})$ 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 $\hat{\mathbf{f}}_{\text{MAP}}$ of $\mathbf{f}$.

   (c) Construct a Laplace approximation to the posterior $p(\mathbf{f}|\mathbf{y})$ and compute the resulting approximation to the posterior predictive $p(f(x_\star)|\mathbf{y})$ for a new input $x_\star$. Compare it to the prediction $p(f(x_\star)|\hat{\mathbf{f}}_{\text{MAP}})$, based on the point estimate $\hat{\mathbf{f}}_{\text{MAP}}$ of $\mathbf{f}$. [*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 $k(x, x') = \kappa\left(\frac{x-x'}{\gamma}\right) = \int \exp\left(i\omega^\top (x - x')\right) \lambda(\omega) d\omega$ 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 γ. 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. 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.

7. Consider the variational approach to GP regression, used not because of non-conjugacy but in order to reduce the computational cost. We have a zero-mean GP prior with covariance k on f and its evaluatons of on training inputs $\{x_i\}_{i=1}^n$, given by vector $\mathbf{f} = [f(x_1), \ldots, f(x_n)]^\top \in \mathbb{R}^n$. We take a small set of inducing inputs $\{z_j\}_{j=1}^m$ and the evaluations of f at these inputs, giving the vector $\mathbf{u} = [f(z_1), \ldots, f(z_m)]^\top \in \mathbb{R}^m$. We then place a variational distribution $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mu, \Sigma)$, which serves as an approximation to the posterior $p(\mathbf{u}|\mathbf{y})$ at these inducing points. On the augmented space $(\mathbf{u}, \mathbf{f})$, we use a variational distribution

$$q(\mathbf{u}, \mathbf{f}) = q(\mathbf{u})\, p(\mathbf{f}|\mathbf{u}),$$

with the true conditional $p(\mathbf{f}|\mathbf{u}) = \mathcal{N}\left(\mathbf{f}|\mathbf{K_{xz}}\mathbf{K_{zz}}^{-1}\mathbf{u}, \mathbf{K_{xx}} - \mathbf{Q_{xx}}\right)$, where $\mathbf{Q_{xx}} := \mathbf{K_{xz}}\mathbf{K_{zz}}^{-1}\mathbf{K_{zx}}$.

   (a) Derive the resulting variational approximation to the posterior $p(\mathbf{f}|\mathbf{y})$ at the training points.

   (b) Prove that

$$\int p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})\, d\mathbf{f} = \log \mathcal{N}\left(\mathbf{y}|\mathbf{K_{xz}}\mathbf{K_{zz}}^{-1}\mathbf{u}, \sigma^2 I\right) - \frac{1}{2\sigma^2} \mathrm{Tr}\left\{\mathbf{K_{xx}} - \mathbf{Q_{xx}}\right\}.$$

   (c) Insert the expression derived in (b) into ELBO, and show that ELBO is maximized for $q(\mathbf{u}) \propto \mathcal{N}\left(\mathbf{y}|\mathbf{K_{xz}}\mathbf{K_{zz}}^{-1}\mathbf{u}, \sigma^2 I\right) p(\mathbf{u})$. Find the value of ELBO for this choice of $q(\mathbf{u})$.

   (d) Compare the derived expression to the exact marginal log-likelihood in the approximate kernel model, which uses the low-rank Nyström approximation $\mathbf{Q_{xx}} = \mathbf{K_{xz}}\mathbf{K_{zz}}^{-1}\mathbf{K_{zx}}$ of $\mathbf{K_{xx}}$.