Problem Set Solutions

1. **Optimization of Marginal Likelihood**

   The optimization of the marginal likelihood w.r.t. the hyperparameters is generally not possible in closed form. Consider, however, the situation where one hyperparameter, θ_0 gives the overall scale of the covariance

   $$k_y(\mathbf{x}, \mathbf{x}') = \theta_0 \tilde{k}_y(\mathbf{x}, \mathbf{x}'),$$

   where k_y is the covariance function for the noisy targets (i.e. including noise contributions) and $\tilde{k}_y(\mathbf{x}, \mathbf{x}')$ may depend on further hyperparameters, $\theta_1, \theta_2, \ldots$. Show that the marginal likelihood can be optimized w.r.t. θ_0 in closed form.

2. **Brownian motion and Brownian Bridge**

   The Brownian motion is a Gaussian process defined for $x \geq 0$ and has $f(0) = 0$. It has mean zero and a non-stationary covariance function $k(x, x') = \min(x, x')$. If we condition it on passing through $f(1) = 0$ we obtain a process known as the Brownian bridge (or *tied-down* Wiener process). Show that this process has covariance $k(x, x') = \min(x, x') - xx'$ for $0 \leq x, x' \leq 1$ and mean 0.

3. **Monotonicity of Predictive Variance**

   Let $\mathrm{Var}_n(f(\mathbf{x}_*))$ be the predictive variance of a Gaussian process regression model at $\mathbf{x}_*$ given a dataset of size n.

   $$\mathrm{Var}_n(f(\mathbf{x}_*)) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{k}_*, \qquad \mathbf{k}_* = \begin{bmatrix} k(\mathbf{x}_*, \mathbf{x}_1) \\ k(\mathbf{x}_*, \mathbf{x}_2) \\ \vdots \\ k(\mathbf{x}_*, \mathbf{x}_n) \end{bmatrix},$$

   and K is the kernel covariance matrix of the observed indices. The corresponding predictive variance using a dataset of only the first $n-1$ training points is denoted $\mathrm{Var}_{n-1}(f(\mathbf{x}_*))$. Show that $\mathrm{Var}_n(f(\mathbf{x}_*)) \leq \mathrm{Var}_{n-1}(f(\mathbf{x}_*))$, i.e. that the predictive variance at $\mathbf{x}_*$ cannot increase as more training data is obtained.

4. **From Basis Functions to GP Marginal Likelihood**

   We have seen that a linear-in-the-parameters model with a Gaussian prior on the weights is mathematically equivalent to a Gaussian Process. Consider a model for periodic, smooth functions defined by the following finite basis:

   $$f(x) = w_1 \cos(x) + w_2 \sin(x)$$

where the weights are assigned independent Gaussian priors: $w_1 \sim \mathcal{N}(0, \alpha)$ and $w_2 \sim \mathcal{N}(0, \alpha)$.

(a) Prove that this model induces a Gaussian Process. Determine its exact mean function $m(x)$ and covariance function $k(x, x')$.

(b) Suppose we record a single noisy observation y_1 at an arbitrary input location x_1, where $y_1 = f(x_1) + \epsilon_1$ and $\epsilon_1 \sim \mathcal{N}(0, \sigma_n^2)$. Using the Gaussian conditioning formulas, derive the explicit posterior predictive mean $\mu_{*|y}$ and predictive variance $\Sigma_{*|y}$ for an arbitrary test point x_*.

(c) The marginal likelihood provides a principled way to optimize hyperparameters by balancing the data fit term and the complexity penalty. Formulate the log marginal likelihood for the single observation y_1 from Part (b) as a function of the hyperparameter α. Find the value of $\hat{\alpha} \geq 0$ that maximizes this likelihood (assuming σ_n^2 is fixed). What happens if $y_1^2 \leq \sigma_n^2$?

5. **Minimax lower bound for sparse sequences**

Let $\Theta := \ell_0[s_n]$ as in PS8, we propose to demonstrate, as $n \to \infty$,

$$\inf_T \sup_{\theta \in \ell_0[s_n]} E_\theta \|T(X) - \theta\|^2 \geq 2 s_n \log(n/s_n)(1 + o(1)),$$

where the infimum is taken over all possible estimators of θ and $s_n = o(n)$, $\log(n)/s_n = o(1)$.

(a) Show that for Π any prior distribution on the whole $\mathbb{R}^n$,

$$R_M := \inf_T \sup_{\theta \in \Theta} E_\theta \|T(X) - \theta\|^2 \geq \inf_T \int_\Theta E_\theta \|T(X) - \theta\|^2 d\Pi(\theta).$$

We consider the following prior distribution on $\mathbb{R}^n$, where $\alpha_n \in (0, 1)$ and $M_n > 0$ are arbitrary for now,

$$\Pi \sim \bigotimes_{i=1}^n (1 - \alpha_n)\delta_0 + \alpha_n \delta_{M_n}.$$

(b) Show that this is a distribution on the set $\Theta_1 := \{0, M_n\}^n$. Do we have $\Pi(\Theta) = 1$?

(c) Show that if θ is drawn according to Π, the preceding infimum can be restricted to the class $\mathcal{S}$ of estimators taking values in $[-2M_n, 2M_n]^n$ only. One could show that the quadratic risk of any estimator is at least as large as that of its 'projection' onto $[-2M_n, 2M_n]$.

(d) Show that, for $\mathcal{S}$ defined in the previous question,

$$\begin{aligned} R_M &\geq \inf_{T \in \mathcal{S}} \int_{\mathbb{R}^n} E_\theta \|T(X) - \theta\|^2 d\Pi(\theta) - \sup_{T \in \mathcal{S}} \int_{\Theta^c} E_\theta \|T(X) - \theta\|^2 d\Pi(\theta) \\ &\geq \inf_T \int_{\mathbb{R}^n} E_\theta \|T(X) - \theta\|^2 d\Pi(\theta) - 10 n M_n^2 \Pi(\Theta^c). \end{aligned}$$

Let $n\alpha_n = s_n - D_n\sqrt{s_n}$, and $M_n := \sqrt{2\log(1/\alpha_n) - C_n}$, with C_n and D_n two sequences that tend to infinity slowly.

(e) Using the deviation inequality $P(|\mathrm{Bin}(n,\alpha_n) - n\alpha_n| > x) \leq 2e^{-x^2/(2a_n(x))}$, with $a_n(x) = n\alpha_n(1-\alpha_n) + x/3$, show that for $D_n = 4(\log n)^{1/2}$, as $n \to \infty$, $nM_n^2\Pi(\Theta^c) = o(1)$.

(f) Express $\inf_T \int_{\mathbb{R}^n} E_\theta\|T(X)-\theta\|^2 d\Pi(\theta)$ as a function of the posterior mean $\overline{\theta} = \int \theta d\Pi(\theta \mid X)$ by interpreting it as a Bayesian risk, then calculate $\overline{\theta}$ explicitly.

(g) Show that for all i,

$$\begin{aligned} m_i := \int E_\theta(\overline{\theta}_i - \theta_i)^2 d\Pi(\theta_i) &\geq \Pi(\theta_i = M_n) E_{M_n}(\overline{\theta}_i - M_n)^2 \\ &\geq \Pi(\theta_i = M_n) E_{M_n}[(\overline{\theta}_i - M_n)^2 \mathbf{1}_{|\varepsilon_i| \leq K_n}], \end{aligned}$$

for any sequence $K_n \to \infty$ and deduce that for $C_n \to \infty$ sufficiently slowly, $m_i \geq \alpha_n M_n^2(1 + o(1))$, uniformly in i.

(h) Conclude.