CSE 515T (Spring 2019) Midterm

- There are two ways to hand in this midterm. **Late submissions will not be accepted!** I do not recommend cutting it too close.
  - Physically in class. The due date for this option is **5:30 PM, Wednesday, 24 April (the last day of class).**
  - Electronically on Piazza as a private message to the instructors. The due date for this option is **23:59 CDT Monday, 29 April** (the midnight of the Monday at the start of reading week).
- Please do not discuss the questions with other members of the class.
- Please post any questions as a *private message to the instructors* on Piazza.
- Any corrections will be posted by the instructors on Piazza. This document will also be kept up-to-date on the course webpage and in GitHub.

In this question we will consider a discrete parameter $\theta \in \{1, 2\}$ with a uniform prior distribution: $\Pr(\theta = 1) = \Pr(\theta = 2) = 1/2$.

A series of observations $\{x_1, x_2, \dots\} \in \mathbb{R}$ will be generated *independently and identically distributed* (iid) according to the following distributions conditional on the unknown parameter θ:

$$p(x \mid \theta = 1) = \mathcal{N}(x; 0, 1^2); \qquad p(x \mid \theta = 2) = \mathcal{N}(x; 1, 2^2).$$

1. (a) Plot the probability density function of the marginal prior predictive distribution for the first value $p(x_1)$ over the interval $x_1 \in [-5, 7]$, and report the value of this pdf at $x_1 = 0$ and $x_1 = 1$ to three decimal places.

   (b) Assume we make an observation $x_1 = 0$. What is the posterior $\Pr(\theta \mid x_1)$?

   (c) What is the posterior mean and variance of $p(x_2 \mid x_1 = 0)$?

   (d) Are there any possible value(s) for x_1 that would not change our prior belief about θ, so that $\Pr(\theta \mid x_1) = \Pr(\theta)$? If so, what is it/are they?

   (e) What would be the observation x_1 that would maximize the posterior probability $\Pr(\theta = 1 \mid x_1)$? Report its value to three decimal places.

Consider the following dataset $\mathcal{D} = (\mathbf{x}, \mathbf{y})$:

$$\mathbf{x} = [-1.30, 0.293, -0.974, -1.44, -1.12, -0.151, 1.08, -0.292, 2.36, -0.04];$$
$$\mathbf{y} = [-1.68, 0.563, -0.954, -0.924, -0.905, -0.833, 0.952, -0.124, 0.838, 0.198].$$

We are going to perform regression with these data assuming the observation model $y = f(x) + \varepsilon$, where $f\colon \mathbb{R} \to \mathbb{R}$ is a latent function we wish to estimate. Assume that the noise ε is generated iid with distribution $p(\varepsilon) = \mathcal{N}(\varepsilon; 0, {}^{1}\!/\!{}_{3^2})$.

Consider two Gaussian process models for f with zero prior mean. Model $\mathcal{M}_1$ has a linear covariance:

$$p(f \mid \mathcal{M}_1) = \mathcal{GP}(f; 0, K_1); \qquad K_1(x, x') = x^\top x'.$$

Model $\mathcal{M}_2$ has a squared exponential covariance with unit length scale and output scale:

$$p(f \mid \mathcal{M}_2) = \mathcal{GP}(f; 0, K_2); \qquad K_2(x, x') = \exp\bigl(-\tfrac{1}{2}(x - x')^2\bigr).$$

2. (a) What is the log model evidence for each model, $\log p(\mathbf{y} \mid \mathbf{x}, \mathcal{M}_i)$?
   (b) What is the model posterior, $\Pr(\mathcal{M} \mid \mathcal{D})$? Report each value to three decimal places.
   (c) Can you find a kernel with higher model evidence given the data above, keeping the mean function and noise variance fixed? (I will award an extra credit point to the person who provides the kernel with the highest evidence.)

In the following assume the model posterior is actually uniform $\Pr(\mathcal{M}_1 \mid \mathcal{D}) = \Pr(\mathcal{M}_2 \mid \mathcal{D}) = {}^{1}\!/\!{}_{2}$, even though it is not.

3. (a) Plot the pdf of the predictive distribution at $x = 4$, $p\bigl(f(x) \mid, x = 4, \mathcal{D}\bigr)$, over the range $-4 \leq f(x) \leq 4$.
   (b) Plot the model-marginal mean function $\mathbb{E}[f \mid \mathcal{D}]$ over the range $-4 \leq x \leq 4$.

We will adopt the second model $\mathcal{M}_2$ *only* for the next question. That is, the prior on f is only

$$p(f) = \mathcal{GP}(f; 0, K); \qquad K(x, x') = \exp\bigl(-\tfrac{1}{2}(x - x')^2\bigr).$$

We will consider a decision problem where we can augment our dataset $\mathcal{D}$ with a single random measurement at a point x^* of our choosing, receiving a measurement $y^* = f(x^*) + \varepsilon$.

Define a loss function $\ell(x^*, y^*, \mathcal{D})$ by the posterior predictive standard deviation of $f(0)$:

$$\ell(x^*, y^*, \mathcal{D}) = \sqrt{\operatorname{var}\bigl[f(x) \mid x = 0, \mathcal{D}, x^*, y^*\bigr]}.$$

4. (a) Plot the expected loss $\mathbb{E}[\ell \mid x^*, \mathcal{D}]$ over the range $-4 \leq x^* \leq 4$. Report its value at $x^* = 1$ to three decimal places.
   (b) What is the optimal final measurement location? Report its value to three decimal places.

Let $f\colon \mathbb{R} \to \mathbb{R}$ have an arbitrary Gaussian process prior:

$$p(f) = \mathcal{GP}(f; \mu, K),$$

where K is differentiable everywhere.

Let $x \in \mathbb{R}$ be an arbitrary point. We will consider inference of the derivative of f at x, $f'(x)$.

Given $h > 0$, consider the value

$$a_h = \frac{f(x+h) - f(x)}{h}.$$

4. (a) What is the variance of a_h? Take the limit as $h \to 0$ and interpret the result.

   (b) Let $x' \in \mathbb{R}$ be another arbitrary point. What is the covariance between a_h and $f(x')$? Take the limit as $h \to 0$ and interpret the result.

   (c) Using the above, find the joint distribution between $f'(x)$ and $f(x')$.

Let $\theta \in \mathbb{R}^d$ be a d-dimensional random variable, and let $p(\theta)$ be an arbitrary prior distribution.

Suppose we discover information $\mathcal{D}$ that informs us every entry of θ is actually positive, without providing any additional information. That is, $\theta_i \geq 0; \forall i$.

5. Describe a rejection sampling procedure for sampling from the posterior distribution $p(\theta \mid \mathcal{D})$.