CSE 515T (Fall 2019) Midterm

- This will be due **11:59 PM Monday, 25 November**. Late submissions will **not** be accepted.
- You will upload your submission to Gradescope.
- 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.

We will consider a series of questions relating to an application of Bayesian inference to numerical analysis, specifically quadrature.

We are going to consider the function

$$f(x) = \exp(-x^2)$$

and its definite integral

$$Z = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x.$$

The function f has no elementary antiderivative, so the calculation of Z is not straightforward. There is a famous method for computing Z with the trick of considering Z^2 instead, rewriting the resulting $2d$ integral in polar coordinates, and making a convenient substitution.[1] If you haven't seen this, it's beautiful and worth checking out. The result is

$$Z = \sqrt{\pi}.$$

We will consider modeling f with a Gaussian process prior distribution:

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

and conditioning on the following set of data $\mathcal{D} = (\mathbf{x}, \mathbf{y})$:

$$\begin{aligned}
\mathbf{x} &= [-2.5, -1.5, -0.5, 0.5, 1.5, 2.5]^\top; \\
\mathbf{y} &= \exp(-\mathbf{x}^2) \\
&= [0.0019305, 0.1054, 0.7788, 0.7788, 0.1054, 0.0019305]^\top.
\end{aligned}$$

We will fix the prior mean function μ to be identically zero; $\mu(x) = 0$.

1. First, let us consider the question of model, specifically kernel, selection.

   Consider the following four choices for the covariance function K:

   $$\begin{aligned}
   K_1(x, x') &= \exp\bigl(-|x - x'|^2\bigr) \\
   K_2(x, x') &= \exp\bigl(-|x - x'|\bigr) \\
   K_3(x, x') &= \bigl(1 + \sqrt{3}|x - x'|\bigr)\exp\bigl(-\sqrt{3}|x - x'|\bigr)
   \end{aligned}$$

   Note that I am not parameterizing any of these kernels; please consider them to be fixed as given.

   Each kernel defines a Gaussian process model for the data in a natural way:

   $$p(f \mid \mathcal{M}_i) = \mathcal{GP}(f; \mu, K_i).$$

   Consider a uniform prior distribution over these models:

   $$\Pr(\mathcal{M}_i) = 1/3 \qquad i = 1, 2, 3.$$

   (a) Compute the log model evidence for each model given the data $\mathcal{D}$ above.

[1] This is commonly credited to Gauss, but the idea goes back at least to Poisson.

(b) Compute the model posterior $\Pr(\mathcal{M} \mid \mathcal{D})$.

(c) Can you find a kernel with higher model evidence given the data above? (I will award an extra credit point to the person who provides the kernel with the highest evidence.)

2. Now let's turn to prediction.

   (a) For each kernel above, plot the predictive distribution over the interval $x^* \in [-6, 6]$. For each model $\mathcal{M}_i$, please plot, in a separate figure, the predictive mean $p(y^* \mid x^*, \mathcal{D}, \mathcal{M}_i)$ and a 95% credible interval. These plots should be the result of a computer program. Please add legends and axes labels, etc., and plot the true function on the same interval for reference. You can take a look through the course materials to get an idea of the sort of plots I am looking for.

   (b) In addition, please write out the predictive mean and standard deviation at $x^* = 0$ for each of the kernels, $p(y^* \mid x^* = 0, \mathcal{D}, \mathcal{M}_i)$.

   (c) What is the model-marginal predictive distribution, $p(y^* \mid x^*, \mathcal{D})$? Write this in terms of the model-conditional predictive distribution and the model posterior.

   (d) Assume that the model posterior is uniform; $\Pr(\mathcal{M}_i \mid \mathcal{D}) = 1/3$ for all models i (this is *not* the case, if you are worried about your answer to 1(b)). Plot the model-marginal predictive mean function $\mathbb{E}[y^* \mid \mathcal{D}]$ over the interval $x^* \in [-6, 6]$.

3. Let us consider a simple numerical estimate of the integral using the midpoint rule. Let $\mathbf{x}_*$ be an evenly spaced grid of n points in the interval $[-6, 6]$ with spacing Δ, starting with $-6 + \Delta/2$ and ending with $6 - \Delta/2$, and let $\mathbf{f}_* = f(\mathbf{x}_*)$. Then a midpoint rule estimate of the integral is

$$\hat{Z}(\Delta) = \Delta \sum_{i=1}^{n} f\big((\mathbf{x}_*)_i\big).$$

   (a) Show that $\hat{Z}(\Delta)$ has a Gaussian distribution. What is its mean and variance?

   (b) Take the limit of our belief about $\hat{Z}(\Delta)$ as $\Delta \to 0$, assuming K can be integrated. Interpret the result in a broader context.

4. Now we will consider integration.

   Perform Bayesian quadrature to estimate the definite integral $\int_{-6}^{6} f(x)\,\mathrm{d}x$, using the model $\mathcal{M}_1$ from question 1. What is the predictive mean and standard deviation, $p(Z \mid \mathcal{D}, \mathcal{M}_1)$? (Please give a numeric answer.) How does this compare with the true answer?

5. Finally, we will consider a decision problem. Suppose we have already made some observations $\mathcal{D}$. How can we select the *most-informative* next observation $\big(x^*, f(x^*)\big)$ to make? This is a decision problem where the action space (parametrizing the next observation location) is the domain, $x^* \in \mathcal{X}$.

   Suppose that we are to estimate Z with a point estimate $\hat{Z}$, and that we have selected the squared loss

$$\ell(Z, \hat{Z}) = (Z - \hat{Z})^2.$$

   (a) Given a set of observations $\mathcal{D}$, what is the Bayesian optimal action? What is the expected loss of that action?

(b) Compute the expected loss of the Bayesian optimal action after adding a new observation to $\mathcal{D}$ located at a point x^*. Plot this result as a function of $x^* \in [-6, 6]$. What is the optimal location to measure the function next? (By symmetry there may be multiple equivalent answers.)

(c) Condition the function on an observation of the function at the chosen location and plot the predictive distribution as in part 2(a). Recompute the predictive distribution for Z. Did our estimate improve?