Question 1:

In the first two problems, we study different approaches to linear regression using a one-dimensional dataset collected from a simulated motorcycle accident. The input variable, \( x \), is the time in milliseconds since impact. The output variable, \( t \), is the recorded head acceleration. The dataset is available on the course webpage. We have divided the full dataset into 40 training examples (variables \( \text{Xtrain} \) and \( \text{Ytrain} \)), and 53 test examples (variables \( \text{Xtest} \) and \( \text{Ytest} \)). To reduce numerical problems, all training features should be scaled to lie in the interval \([-1, +1]\) before fitting. Note that an equivalent scaling must then be applied to all test data. We have provided a demonstration script to get you started.

We compare linear regression models with various families of non-linear basis functions. For a model of order \( L \), we define \( M = L+1 \) basis functions: the constant function \( \phi_0(x) = 1 \), and \( L \) functions \( \phi_j(x), j = 1, \ldots, L \), that vary with \( x \). In your solutions, you do not need to report the numerical values \( \hat{w} \) of estimated regression coefficients. We will assess correctness using the plots you create based on these estimates.

In the standard linear regression model, the observations \( t_n \) follow a Gaussian distribution centered around a linear function \( w \) of a fixed set of basis functions:

\[
p(t_n | x_n, w, \beta) = \text{Normal}(t_n | w^T \phi(x_n), \beta^{-1}).
\]

Here, \( \beta \) is the inverse variance or precision. As derived in the textbook, given \( N \) training observations the maximum likelihood weights \( \hat{w} \) satisfy the normal equations

\[
(\Phi^T \Phi) \hat{w} = \Phi^T t,
\]

where \( t \) is a \( N \)-dimensional vector of training responses \( t_n \), and \( \Phi \) is an \( N \times M \) matrix of corresponding features, such that \( \Phi_{ij} = \phi_j(x_i) \).

a) Use the provided function \texttt{basisPoly} to define a family of polynomial basis functions \( \phi_j(x) = x^j \). On a single set of axes plot, versus \( x \), the test data and the mean prediction \( \hat{w}^T \phi(x) \) for a model of order \( L = 5 \) with parameters \( \hat{w} \) estimated via maximum likelihood.

**Hint:** To compute \( x = A^{-1} b \) in Matlab, rather than explicitly calling the \texttt{inv} command, use the following command to improve numerical stability:

\[
>> \ x = A \ \backslash \ \ b;
\]
b) Compute maximum likelihood (ML) estimates \( \hat{w} \) of the regression parameters for models of order \( L = 0, 1, 2, \ldots, 19 \). Consider the following squared error metric:

\[
L(x, t \mid \hat{w}) = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (t_n - \hat{w}^T \phi(x_n))^2}
\]

Evaluate and plot \( L(x, t \mid \hat{w}) \) as a function of the model order, \( L \), for the 40 training examples. Also do this (on the same axes) for the 53 test examples. Which model has the smallest training error, and which has the smallest test error? Plot the mean predictions \( \hat{w}^T \phi(x) \) for both of these models on the same axes as in part (a).

c) We now consider alternative, radial basis functions of the form

\[
\phi_j(x) = \exp\left(-\frac{(x - \mu_j)^2}{2\sigma^2}\right), \quad j = 1, \ldots, L.
\]

For any model order \( L \), we space the basis function centers \( \mu_j \) evenly between \(-1\) and \(1\), and set the bandwidth \( \sigma = (\mu_2 - \mu_1) \) to the distance between basis centers. See the provided function `basisRadial`. Repeat parts (a-b) for radial basis function models of order \( L = 0, 5, 10, 15, 20, 25 \).

d) In constructing the training and test sets, we excluded one point from the original dataset: \( x = 57.6, t = 10.7 \). What is the error in the prediction of this point for the polynomial model which minimized the loss on the other test points? What is the error in the prediction of this point for the radial basis function model which minimized the loss on the other test points? Discuss any qualitative differences between this datapoint and the other test data.

Question 2:

In the previous question, you may have noticed that unregularized ML estimates can become unstable for large model orders. We now consider an alternative, Bayesian approach in which the regression coefficients are assigned a Gaussian prior

\[
p(w) = \text{Normal}(w \mid 0, \alpha^{-1}I_M)
\]

where \( \alpha > 0 \) is an inverse-variance hyperparameter. As derived in the textbook, the posterior mean \( m_N \) and covariance \( S_N \) of the weight vector, under a Gaussian prior, equal

\[
m_N = (\Phi^T \Phi + \frac{\alpha}{\beta}I_M)^{-1} \Phi^T t, \quad S_N = (\beta \Phi^T \Phi + \alpha I_M)^{-1},
\]

where \( I_M \) is an \( M \times M \) identity matrix.

a) Give an expression for the MAP estimate of \( w \) under the Gaussian prior above, and the linear observation model of question 1. For a polynomial basis of order \( L = 50 \) and \( N = 40 \) training examples, is there a unique MAP estimate? Is there a unique ML estimate? Explain your reasoning.
b) Fix $\beta = 0.0025$, and consider 100 candidate values for the regularization parameter $\alpha$, logarithmically spaced between $10^{-8}$ and $10^0 = 1.0$:

```matlab
>> alpha = logspace(-8,0,100);
```

Determine MAP estimates $\hat{w}$ for each of these priors, and a polynomial basis of order $L = 50$. Using the `semilogx` command, plot the error metric of problem 1(b) versus $\alpha$ for both the training and test datasets. Then plot the mean prediction $\hat{w}^T \phi(x)$ for the models which minimize the training and test error.

c) Repeat part (b) for a radial basis function model of order $L = 50$.

d) Consider the pair of models from parts 2(b,c) which minimize the test error for their corresponding basis families. Describe how the Gaussian posterior distribution on $w$ can be used to find the posterior of the prediction function $f(x) = w^T \phi(x)$, for any fixed input $x$. Draw and plot 10 samples from the polynomial and radial basis function posteriors on prediction functions, for the grid of $x$ values defined in the example code.

Hint: The Matlab `mvnrnd` command samples from a multivariate normal distribution.

e) Again consider the pair of models from part 2(d). What is the error in their corresponding predictions of the held-out test point from problem 1(d)? Are the relative magnitudes of these errors predictable from the posterior distributions plotted in part 2(d)?

Question 3:

In this question, we consider a continuous estimation problem in which the input $x$ and response variable $t$ are both real numbers. Their joint probability density function $p(x,t)$ is constant within the closed, shaded region shown in Figure 1, and zero elsewhere.

a) For any estimator $y(x)$, consider the quadratic loss function $L(y(x), t) = (y(x) - t)^2$. Determine the estimator that minimizes the following posterior expected loss:

$$\mathbb{E}[L(y(x), t) \mid x] = \int_{-\infty}^{\infty} L(y(x), t) p(t \mid x) \, dt$$

Simplify the form of your answer as much as possible.

b) Consider how the estimator $y(x)$ from part (a) behaves for three particular input variables: $x = -1.5$, $x = 0$, and $x = 1.5$. In each case, is $y(x)$ also a maximum a posteriori (MAP) estimator? Why or why not?

c) Suppose that rather than knowing the true density function $p(x,t)$, we instead have $N$ training examples $\{(x_n, t_n)\}_{n=1}^N$, where $(x_n, t_n)$ are independent samples from $p(x,t)$. Consider the following linear regression model:

$$p(t \mid w, x, \beta) = \text{Normal}(t \mid w^T \phi(x), \beta^{-1})$$

Suppose that we choose $\phi(x) = [1, x]^T$ as our basis or feature functions, and estimate $\hat{w}$ via maximum likelihood. As $N \to \infty$, will the expected loss of the prediction function $y_n(x) = \hat{w}^T \phi(x)$ be lower than, higher than, or equal to that of the estimator from part (a)? Justify your answer.
Figure 1: Joint probability density function $p(x,t)$ for question 3. The density is constant (uniform) in the shaded region, and zero elsewhere.

d) Consider again the linear regression model from part (c), but with alternative features $\phi'(x) = [1, |x|^2]^T$. As $N \to \infty$, will the expected loss of $y'_\theta(x) = \hat{\omega}^T \phi'(x)$ be lower than, higher than, or equal to that of the estimator from part (a)? Justify your answer.