Problem:

In this question you will implement linear regression with Gaussian bases. Start by downloading the dataset from Avenue. The dataset is the AutoMPG dataset from the UCI repository. The task is to predict fuel efficiency (miles per gallon) from 7 features describing a car. Load the data and normalizing the features and targets to have 0 mean and unit variance if necessary. Implement linear basis function regression with Gaussian basis functions. For the centers μ_j use randomly chosen training data points. Set covariance matrix $\Sigma = 2I$. Perform the following experiments:

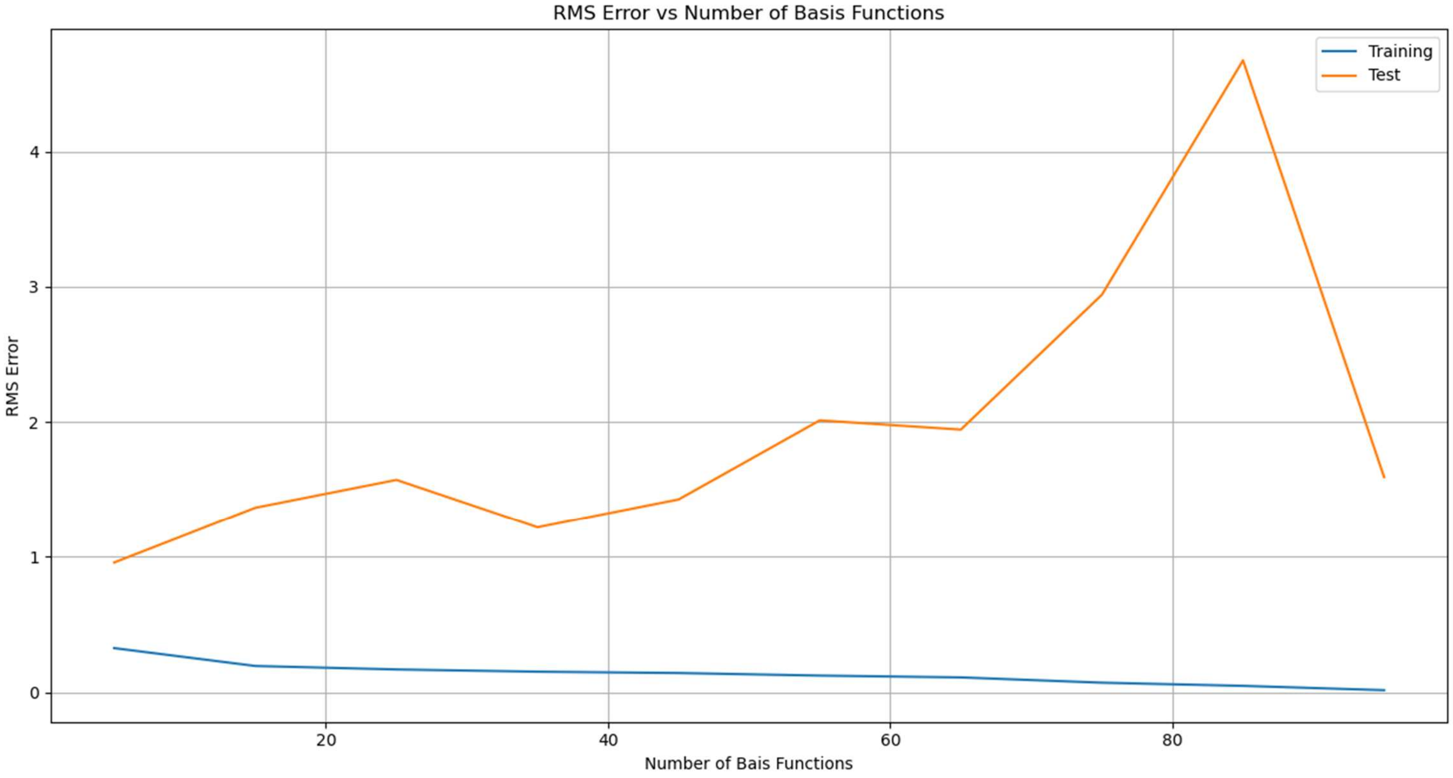
a) Using the first 100 points as training data and the remainder as testing data. Fit a Gaussian basis function regression using 5, 15, 25, . . ., 95 basis functions. Do not use any regularization. Plot training error and test error (in RMS error) versus number of basis-functions. Put this plot, along with a brief comment on what you see, in your report.

b) Implement L2-regularized regression. Again, use the first 100 points. Fit a regression model with 90 basis functions using λ= {0, 0.01, 0.1, 1, 10, 100, 1000}. Use 10-fold cross-validation to decide on the best

value for λ. Produce a plot of validation set error versus regularizer value. Use a semilogx plot, putting regularizer value on a log scale. Put this plot in your report and note which regularizer value you would choose from the cross-validation.

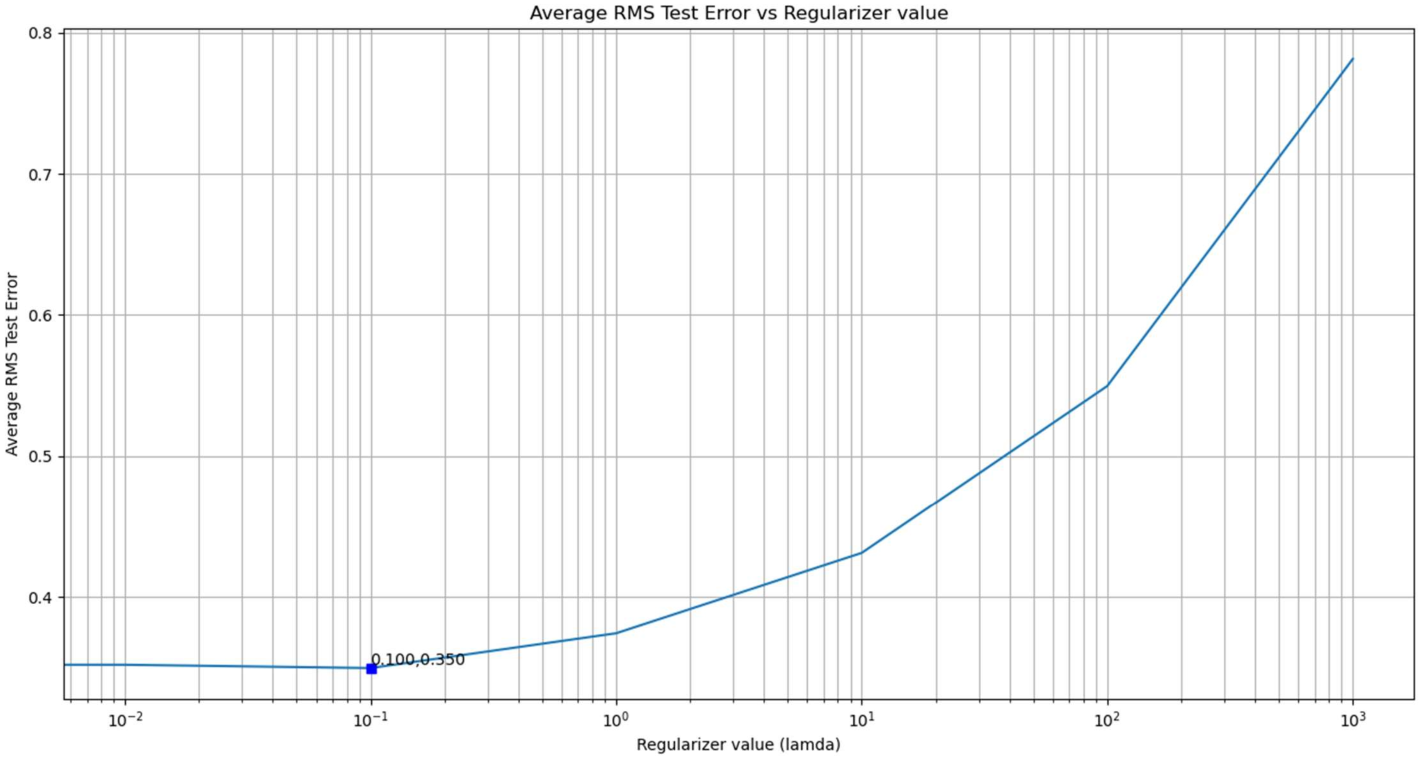
Solution:

a) **question1a.py** is the python script for this question



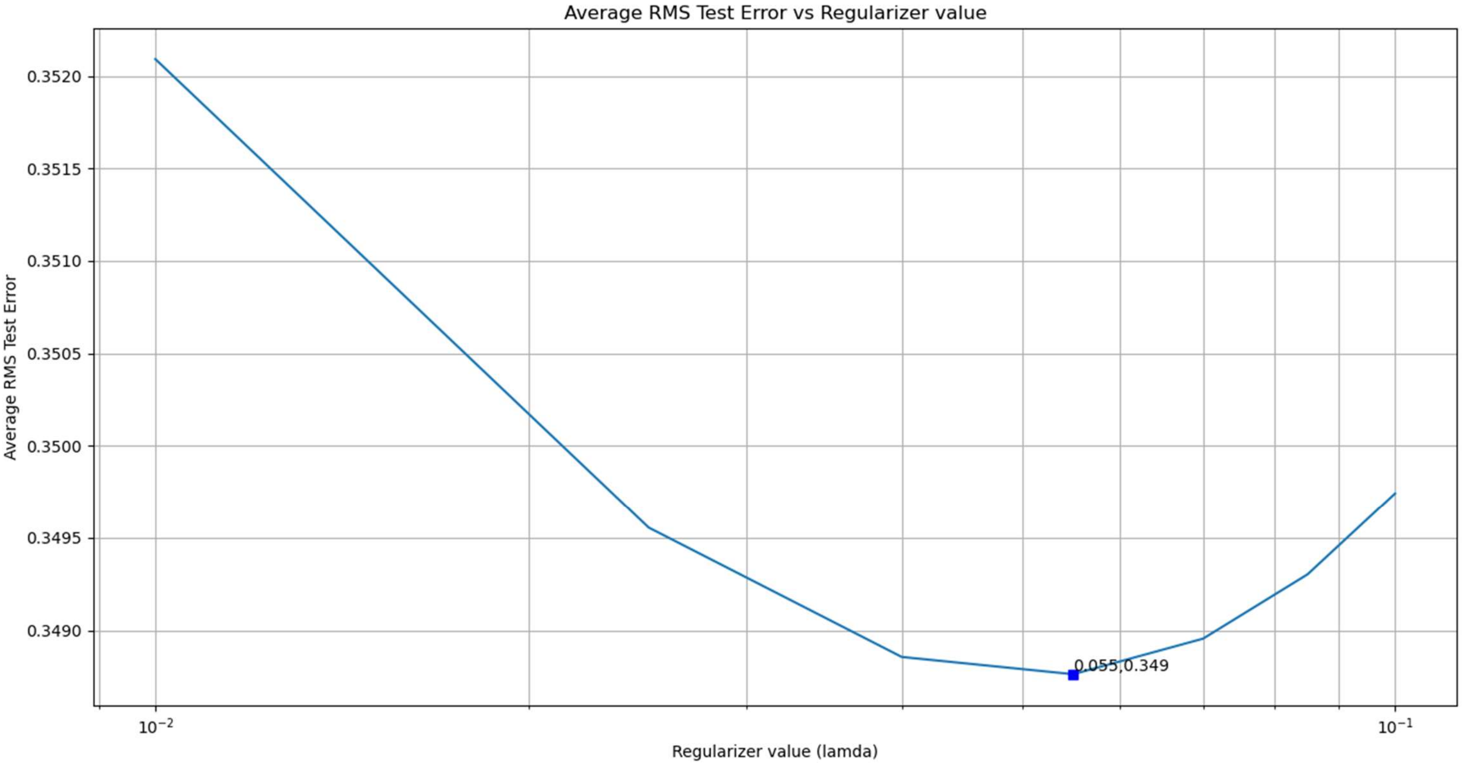
The plot above indicates that with an increase in basis functions we get an increase in test error and decrease in training error. The increase in test error is attributed to overfitting the data as we are creating a model with increasing complexity with more coefficients with very high values. The decrease in training error is a result of the complex model being a very good fit for the training data, however, a terrible fit for any test values.

b) **question1b.py** is the python script for this question



Seeing as how the plot from part (a) was clearly overfitting the dataset, L2 regularization was added to the regression model to get a better fit.

The plot above was generated by doing 10-fold cross validation to determine the best value for lamda, the regularizer value in the L2 regularization. There are values of lamda being tested (0, 0.01, 0.1, 1, 10, 100, 1000). From the plot above we can see that we get a minimum RMS test error for lamda = 0.1. To be more precise on the most ideal regularizer value I have gone one step further and defined a new range of values for lamda. This new range is defined by lamda = np.linspace(0.01,0.1,7), which are seven values from 0.01 to 0.1. The plot for this new range is provided below.



From this new plot we can see the minimum average RMS test error occurs for a lamda of 0.055. This exercise can be repeated to get a more precise lamda value, however, seeing as how the minimum average RMS test error is not reducing by a significant amount I will stop at this point.