Question 1:
You are to implement regularized polynomial curve fitting. The examples are \((x_n, t_n), n = 1..N\). \(\omega_i\) is the weight for \(x_n^i\) (0 <= i <= 9).

\[
\omega^* = \arg\min_{\omega} \left( \sum_n \left( \sum_i \omega_i x_n^i - t_n \right)^2 + \lambda \sum_i \omega_i^2 \right)
\]

Let’s rewrite the above in matrix notation: \(X\) is a \(10 \times N\) matrix, where \(N\) is the number of examples and the \(n\)th column of \(X\) is \((1, x_n, x_n^2, ..., x_n^9)^T\). \(t\) is the vector of outputs \(t=(t_1, t_2, ..., t_n)^T\).

\[
\omega^* = \arg\min_{\omega} \left( \|X^T \omega - t\|^2 + \lambda \|\omega\|^2 \right)
\]

In our dataset, each row contains one value of \(x\) and the corresponding value of \(t\), separated by space. Use 10-fold cross validation to find the best choice of \(\lambda\) and report the loss on the test set.

Extra credit:
• Modify the above solution so that the bias term / weight is not regularized. Justify your answer.
• Show variance bars for the losses of the different holdouts.
• Implement leave one out cross validation and show that in some sense this is better than 10-fold cross validation. Remember you to look at the test score only once to report results!

Solution:
Finding the Optimal \(\lambda\):
For calculating the minimum \(\omega\), we need to calculate the derivative of \(\|X^T \omega - t\|^2 + \lambda \|\omega\|^2\) and then get the minimum \(\omega\) through making the derivative equals 0. The process of getting the derivation is as following:

\[
X(X^T \omega^* - t) + \lambda w^* = 0
\]
\[
(XX^T + \lambda I)w^* - Xt = 0
\]
\[
w^* = (XX^T + \lambda I)^{-1} Xt
\]

Next we compare the performance of 10-fold cross validation and leave-one-out cross validation (The code is almost the same as the 10-fold cross validation, in fact, leave-one-out is 20-fold cross validation in this homework because we have 20 data points). We tried 12 different lambdas (0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10). The following are the experimental results of loocv and 10-fold. The x axis is \(\lambda\) and the y axis is 5000 times mean square errors (Because we want to show the results more clearly).

<table>
<thead>
<tr>
<th>Bias Regularization</th>
<th>CV Method</th>
<th>10-Fold</th>
<th>LOOCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal lambda</td>
<td>5.0</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Test Errors</td>
<td>0.817</td>
<td>0.678</td>
<td></td>
</tr>
</tbody>
</table>
And we set the first element of Identify Matrix to 0 and modify the above solution so that the bias term / weight is not regularized. The following are the experimental results of loocv and 10-fold. The x axis is $\lambda$ and the y axis is 5000 times mean square errors (Because we want to show the results more clearly).
Without Bias Regularization

<table>
<thead>
<tr>
<th>CV Method</th>
<th>10-Fold</th>
<th>LOOCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal lambda</td>
<td>0.2</td>
<td>0.05</td>
</tr>
<tr>
<td>Test Errors</td>
<td>0.0898</td>
<td>0.0724</td>
</tr>
</tbody>
</table>

10-fold Cross Validation without Bias Regularization

Leave-one-out Cross Validation without Bias Regularization