Q1.

The 2X2 matrix of mean square error for the training and test for 100 and 10 training test points is shown below.

From the below table it is pretty clear that the training and the test error depend on the training set size.

The training set error for 10 data point is less as it can train better on the small data set, where as on the 100 data points it has a slightly higher error as there are more points to train over.

The test error is slightly higher on the 10 data points as the trained model calculated from the training set is more generalised to the variation in the training set, as compared to the 100 training points, where the model is better generalised as compared to the 10 training points.

Because of this change in how the model is getting trained the test error for higher number of training set is lesser than that on the smaller set.

	10 point training	100 point training
Training set	0.9039	0.9898
Test set	1.1204	1.0052

Fig 1 : figure of the 2x2 matrix for training and test error on 10 and 100 data points

Q2.

The 2X2 matrix of mean square error for the training and test for 100 and 10 training test points over 10 dimensions is shown below.

From the below table it is pretty clear that the training and the test error depend on the training set size and dimension.

The training set error for 10 data point on the 10 dimensions is almost zero. As the model was able to fit to the small data provided almost perfectly. Where as for a larger set the training error is more as there are more number of variations in the data and the model was not able to generalise As well as it did on a smaller dataset

The test error is very high for the 10 datapoint model as the model has overfit over the training set and is not able to predict better. Where as for the 100 datapoint model the test error is pretty close compared to the 10 point error as the model was able to generalise better over the training set better and was able to do better predictions over the test set.

Because of this change in how the model is getting trained the test error for higher number of training set is lesser than that on the smaller set.

	10 point training	100 point training
Training set	-8.7930e-17	0.8904
Test set	1.2162e+03	1.1190

Fig 2 : figure of the 2x2 matrix for training and test error on 10 dim data

Q3.

1. $w^* = \underset{w}{\operatorname{argmin}}\ \gamma w^T w + \frac{1}{L}\sum_{i=1}^{L}(x_i^T w - y_i)^2$

$\frac{\partial}{\partial w^*}\left(\gamma w^{*T} w^* + \frac{1}{L}\sum_{i=1}^{L}(x_i^T w^* - y_i)^2\right) = 0$

$\Rightarrow \gamma \frac{\partial}{\partial w^*} tr(w^{*T} w^*) + \frac{1}{L}\frac{\partial}{\partial w^*}(Xw^* - y)^T(Xw^* - y) = 0$ (Where $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_L^T \end{bmatrix}$)

$\Rightarrow \gamma(2w^*) + \frac{1}{L}\frac{\partial}{\partial w^*}(w^{*T}X^TXw^* - y^TXw^* - w^{*T}X^Ty + y^Ty) = 0$ (Using $\frac{\partial}{\partial A} tr(A^TBAC) = BAC + B^TAC^T$)

$\Rightarrow 2\gamma w^* + \frac{1}{L}\frac{\partial}{\partial w^*}(tr(w^{*T}X^TXw^*) - 2tr(w^{*T}X^Ty) + y^Ty) = 0$ (All the terms inside are scalars)

$\Rightarrow 2\gamma w^* + \frac{1}{L}((X^TXw^* + X^TXw^*) - 2X^Ty) = 0$

$\Rightarrow 2\gamma w^* + \frac{1}{L}(2X^TXw^* - 2X^Ty) = 0$

$\Rightarrow \gamma L w^* + X^TXw^* = X^Ty$

$\Rightarrow X^TXw^* + \gamma L w^* = X^Ty$

$\Rightarrow (X^TX + \gamma L I)w^* = X^Ty$

$\Rightarrow w^* = (X^TX + \gamma L I)^{-1}X^Ty$ (This is well-defined since $X^TX + \gamma L I$ is positive definite by part 2 (to follow) and all positive definite matrices are invertible)

2. Let v be a non zero vector:

$v^T(X^TX + \gamma L I)v = v^TX^TXv + \gamma L v^TIv$

$= (Xv)^T(Xv) + \gamma L v^Tv$

$= \sum_{i,j}(x_{ij}v_j)^2 + \gamma L \sum_j v_j^2$

$\geq \gamma L \sum_j v_j^2$

Assuming we have at least 1 data point and that the regularization parameter is greater than zero, we have: $\gamma > 0$, $L > 0$ and $v_j > 0$ for at least one value of j since $v \neq 0$.

So $\gamma L \sum_j v_j^2 > 0$

$\therefore X^TX + \gamma L I$ is positive definite.

Fig 3 : the above figure gives the solution for part a and b of question 3

Q4.

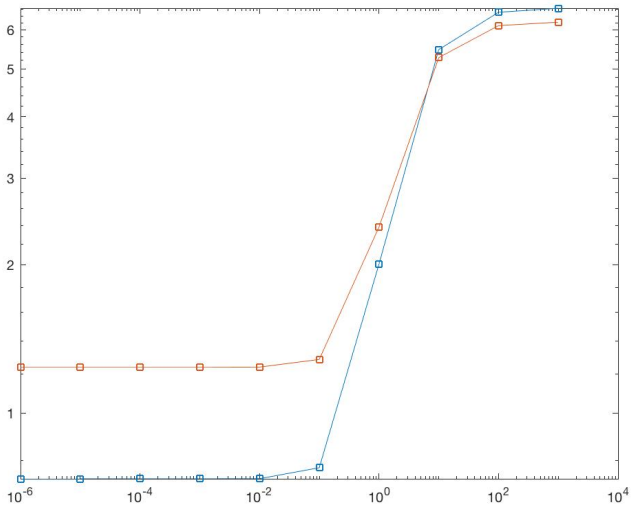

Fig 4.1: train & test error on 100 training points

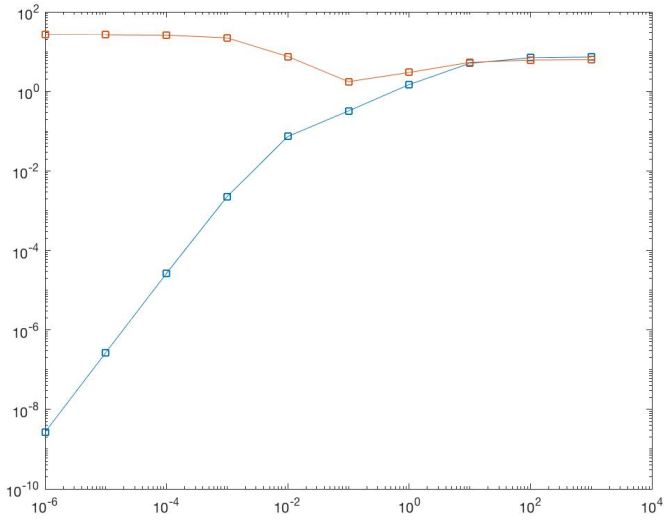

Fig 4.2: train and test error on 10 training points

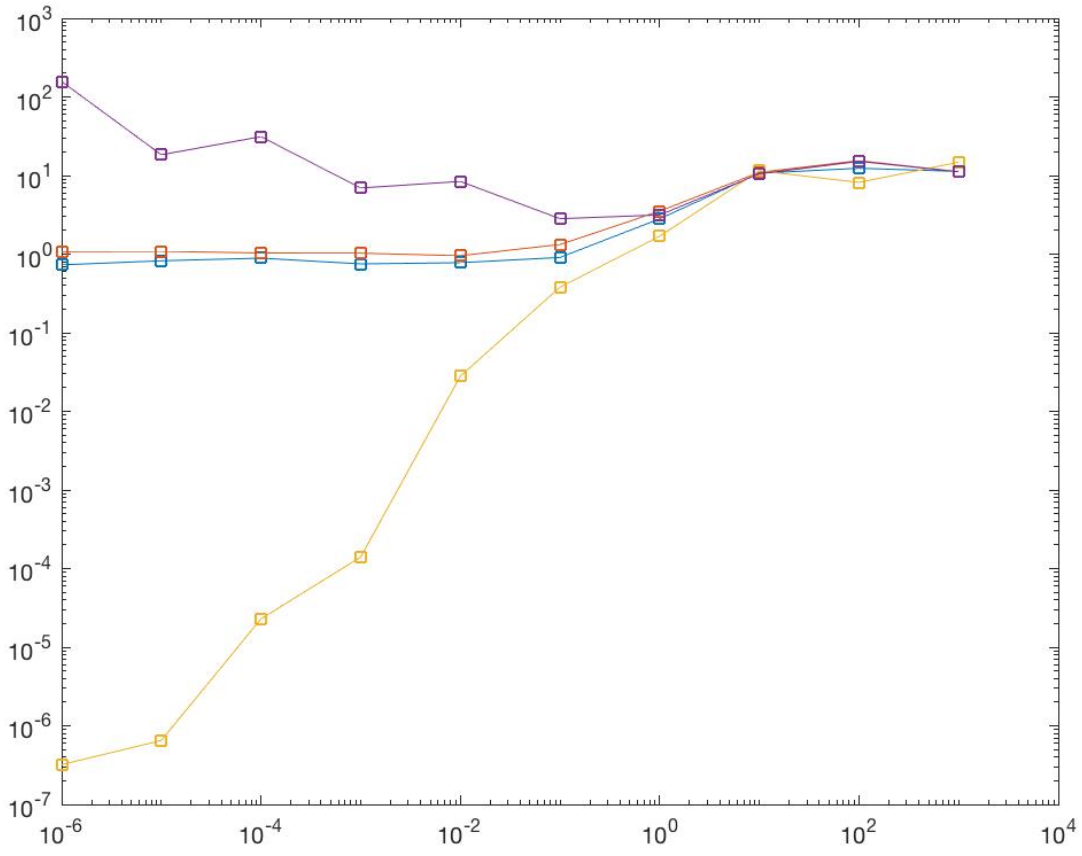

Fig 4.3 : train and test errors calculated 200 times and averaged over 10 and 100 datapoint

All the above images are plotted on a log scale.

From the above graphs it is very clear that the least test error can be achieved by choosing the least train error achieved from a range of regularisation parameter γ.

In the above example we have chosen a range of 10^{-6} to 10^{+3}. And we achieved a test and train errors as shown below for the range of regularisation parameter mentioned above.

	Training Error	Test Error
1	0.7354	1.2381
2	0.7354	1.2381
3	0.7354	1.2381
4	0.7354	1.2382
5	0.7359	1.2398
6	0.7745	1.2848
7	2.0077	2.3871
8	5.4699	5.2691
9	6.5124	6.1183
10	6.6357	6.2184

Fig 4.4 : training and test error on 100 points

	Training Error	Test Error
1	2.6859e-09	27.5642
2	2.6815e-07	27.5178
3	2.6375e-05	27.0598
4	0.0023	23.0589
5	0.0733	7.7925
6	0.3221	1.8409
7	1.4810	3.0119
8	5.0840	5.2989
9	7.0201	6.1082
10	7.3046	6.2172

fig 4.5 : training and test error on 10 points

	Training Error	Test Error	Training Error 10 pts	Test Error 10 pts
1	0.7306	1.0751	3.2057e-07	158.1250
2	0.8226	1.0863	6.5113e-07	17.9705
3	0.8893	1.0343	2.2909e-05	32.1677
4	0.7513	1.0282	1.4005e-04	6.9902
5	0.7779	0.9546	0.0280	8.1223
6	0.9072	1.3267	0.3796	2.8750
7	2.8116	3.5222	1.6879	3.1842
8	10.7081	10.9785	11.4841	10.5564
9	12.3664	15.4502	8.1663	15.5466
10	11.2767	11.2012	14.7555	11.1925

Fig 4.6 : train and test error for 100 and 10 points averaged over 200 times

From the above 3 figures it is clear that as the number of data points increase the lower the train error the lower is the test error. But if we take a lower number of data points there is still an issue of overfitting but not as drastic as the previous method. The overfit test error is far less compared to previous results

Q5.

a/b

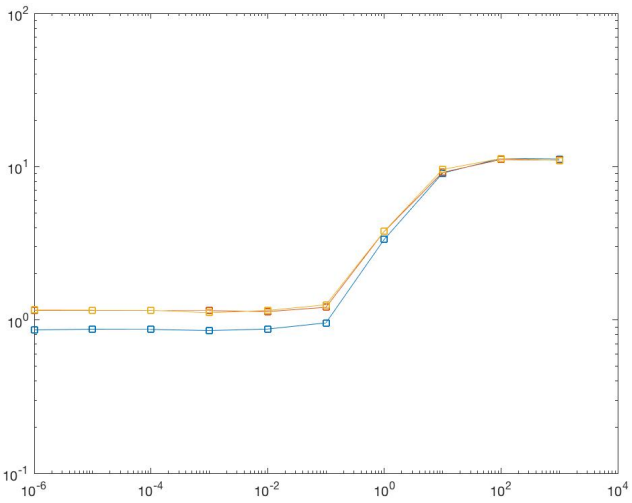

Fig 5.1: train, test, val errors for 100 data points

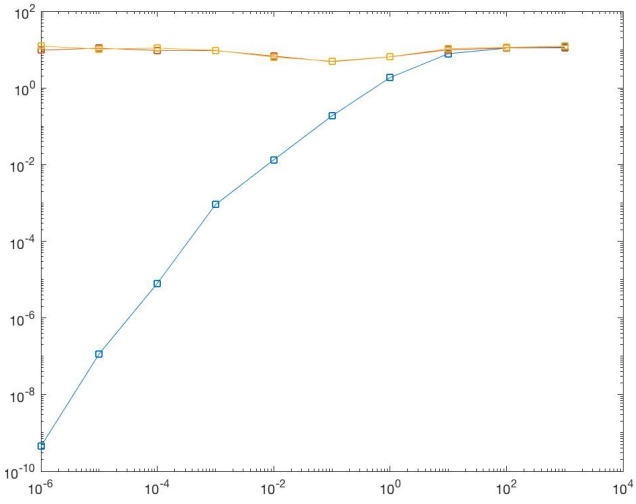

Fig 5.2 : train, test, val error for 10 data points

From the above graphs the min validation error for both 100 and 10 training sets were chosen and the corresponding regularisation parameter was selected and the following test errors were calculated.

	Test Error
100 points	1.1144
10 points	3.8516

The validation errors were pretty close to the test errors so selecting the least validation error helped us to train the model to get the lowest possible test error

c.

	Mean Gamma
100 points	0.0383
10 points	102.0421

From the above table we can see that the larger dataset like the 100 training points with 10 dimensions has a lower regularisation parameter as compared to the 10 dimensional 10 data points

This is because the 10 data point training model will overfit and have a high validation error in the beginning and as the regularisation parameter is increased the validation error reduces to a minimum value and hence the test error, so they have a high regularisation parameter

In the 100 data point 10 dimensional training, the lower regularisation parameter leads to a lower validation error hence lower mean regularisation parameter

d.

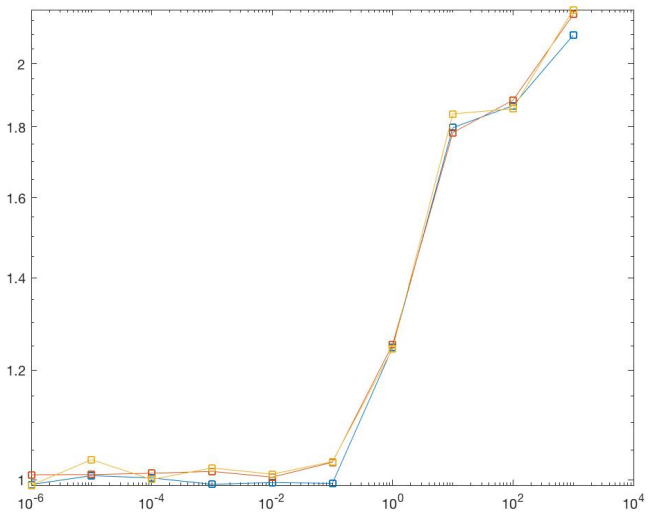

Fig 5.1: train, test, val errors for 1D-100 data points.

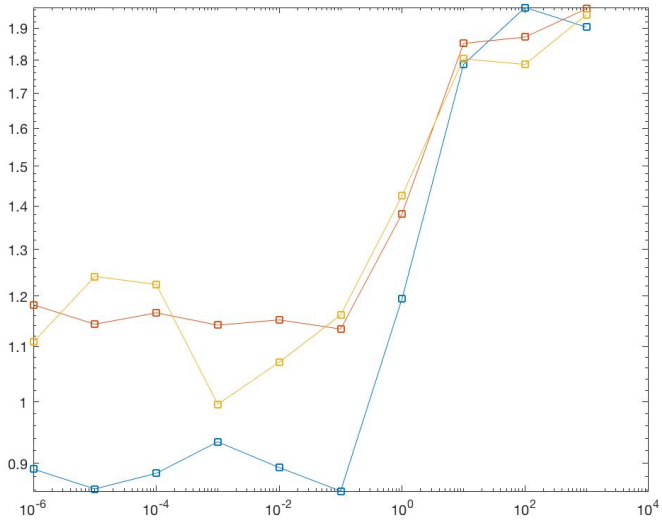

Fig 5.2 : train, test, val error for 1D-10 Data points

	Test Error
100 points	1.0064
10 points	1.1296

Similar to problem a/b, the validation errors were pretty close to the test errors so selecting the least validation error helped us to train the model to get the lowest possible test error

Q6.

	Training Error	Val Error	Test Error
1	0.8526	1.1118	1.1564
2	0.8586	1.1182	1.1497
3	0.8610	1.1280	1.1509
4	0.8598	1.1248	1.1500
5	0.8614	1.1268	1.1458
6	0.8648	1.1294	1.1425
7	0.8642	1.1282	1.1423
8	0.8681	1.1317	1.1427
9	0.8691	1.1327	1.1432
10	0.8708	1.1339	1.1409

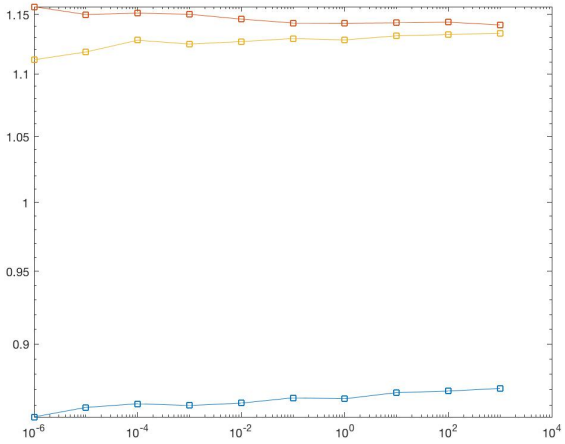

Fig 6.1 table for 100 data points after cross validation. Along with the graph for different values of gamma

	Training Error	Val Error	Test Error
1	0.8473	6.1214	6.4218
2	0.8069	6.8969	6.5120
3	0.7650	6.4110	6.3593
4	0.7251	6.2317	6.3542
5	0.6882	6.2660	6.4062
6	0.6543	6.3200	6.4669
7	0.6234	6.3068	6.4306
8	0.5952	6.3153	6.4703
9	0.5694	6.3445	6.5110
10	0.5457	6.4186	6.5162

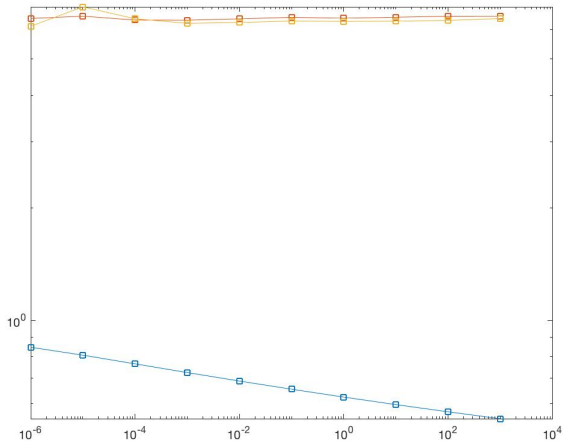

Fig 6.2 table for 10 data points after cross validation. Along with the graph for different values of gamma

Q7.

a. For 100 data points the mean test error, mean train error and the standard deviation of the train error are shown below

	Ex4	Ex5	Ex6
Mean Train Error	4.0451	4.3028	1.1517
Std train error	4.5806	4.5510	0.1090
Mean test error	1.1143	1.1478	1.1182

b. For 10 data points the mean test error, mean train error and the standard deviation of the train error are shown below

	Ex4	Ex5	Ex6
Mean Train Error	3.2524	9.5526	3.7011
Std train error	4.7577	2.6504	1.8197
Mean test error	527.0835	4.8341	10.2305

Q9.

	train errror	+/-std	test error	+/-std
Naive Regression	88.0139	8.7793	84.2354	11.2616
attribute 1	71.8022	7.0551	73.9653	0
attribute 2	73.9305	7.1411	77.6448	2.9160e-14
attribute 3	64.3422	7.9670	61.2259	0
attribute 4	81.8559	6.8926	86.1508	2.9160e-14
attribute 5	69.8259	7.7610	69.8310	1.4580e-14
attribute 6	43.5553	6.7956	47.8494	1.4580e-14
attribute 7	73.3069	8.0142	75.4079	2.9160e-14
attribute 8	79.9279	7.8106	79.3997	0
attribute 9	72.4566	8.1110	73.3307	1.4580e-14
attribute 10	66.0186	7.9813	66.0362	0
attribute 11	61.7094	6.9255	66.2659	1.4580e-14
attribute 12	75.5050	6.5230	73.1373	0
attribute 13	38.3912	3.4881	38.0622	7.2900e-15
attribute all	19.7507	2.9713	23.6997	7.2900e-15