University of Michigan

EECS 545: Homework 2

Zhijie Dong

February 12, 2018

1 Problem 1

1.1 Part A

In this subsection, we reuse the closed form, the predicted labels is plotted in Fig 1.1

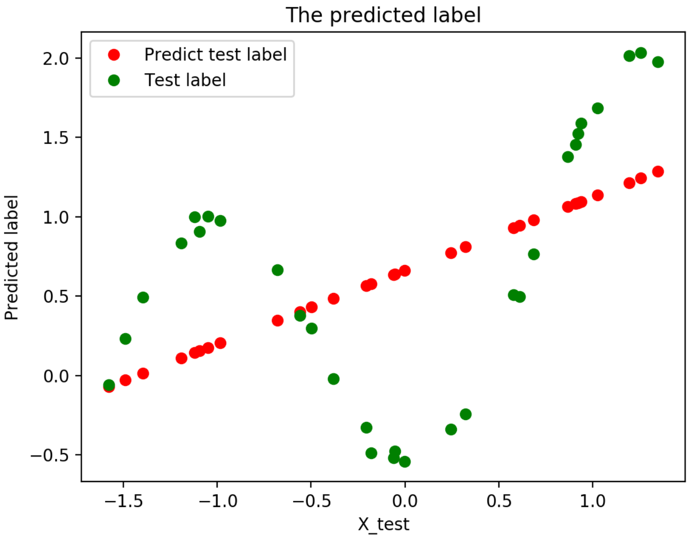
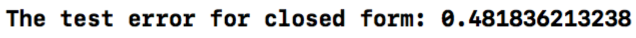


Figure 1.1: Predicted labels for closed form

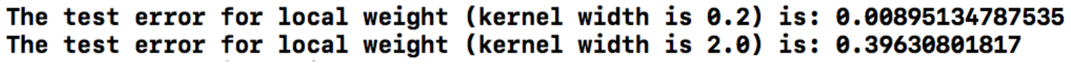
The test error is:



```
The test error for closed form: 0.481836213238
```

1.2 Part b

In this subsection, we fit a locally weighted linear regression model for different sigma, the corresponding plots of predicted label are shown in Fig 1.2 and Fig 1.3, and test errors are:



```
The test error for local weight (kernel width is 0.2) is: 0.00895134787535
The test error for local weight (kernel width is 2.0) is: 0.39630801817
```

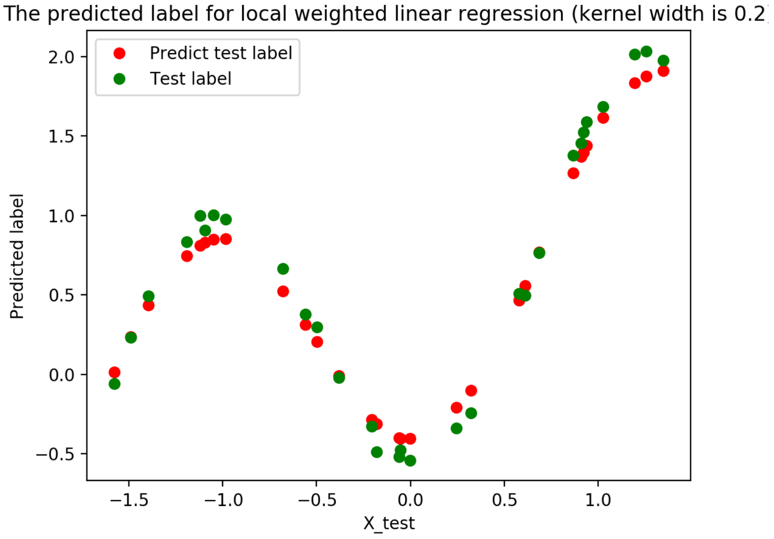


Figure 1.2: Predicted labels for locally weighted linear regression model ($\tau = 0.2$)

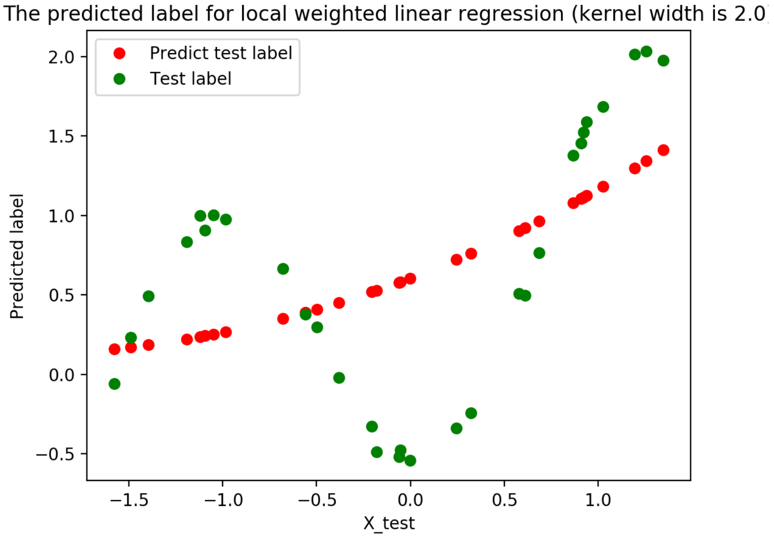
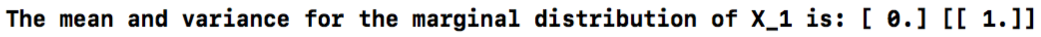


Figure 1.3: Predicted labels for locally weighted linear regression model ($\tau = 0.2$)

2 Problem 2

2.1 Part B

The marginal distribution of X_1 is:



```
The mean and variance for the marginal distribution of X_1 is: [ 0.] [[ 1.]]
```

And the distribution is plotted in Fig 2.1.

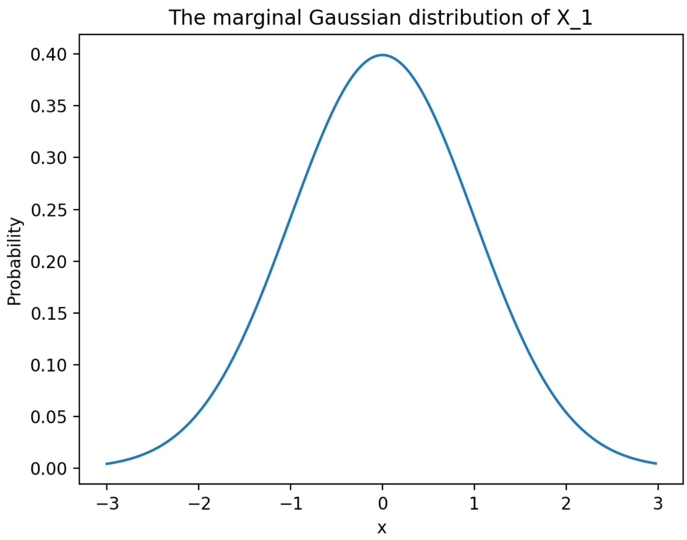
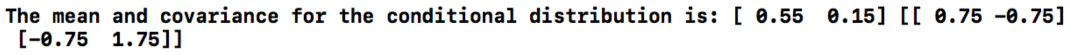


Figure 2.1: Marginal distribution of X_1

2.2 Part D

The conditional distribution $P(X_1, X_4; X_2 = 0.1, X_3 = -0.2)$ are computed as follows, and the distribution is plotted in Fig 2.2 and Fig 2.3.



```
The mean and covariance for the conditional distribution is: [ 0.55  0.15] [[ 0.75 -0.75]
 [-0.75  1.75]]
```

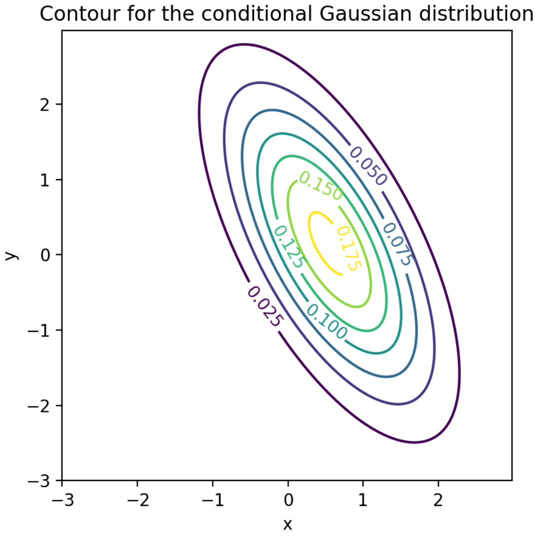


Figure 2.2: Contour of distribution $P(X_1, X_4; X_2 = 0.1, X_3 = -0.2)$

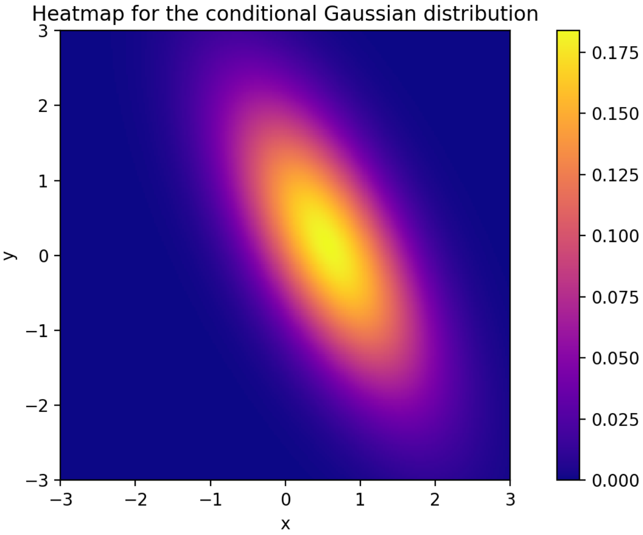
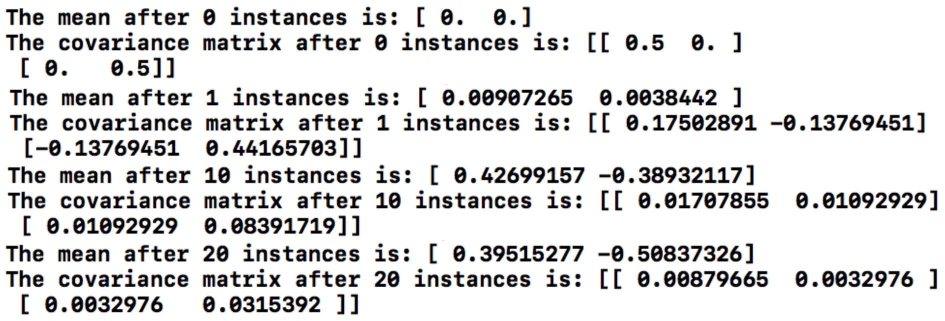


Figure 2.3: Heat map of distribution $P(X_1, X_4; X_2 = 0.1, X_3 = -0.2)$

3 Problem 3

The initial distribution of w and the distribution after 1, 10, 20 instances are shown as follows, and corresponding heatmaps are shown in Fig 3.1, Fig 3.2, Fig 3.3 and Fig **??**.



```
The mean after 0 instances is: [ 0.  0.]
The covariance matrix after 0 instances is: [[ 0.5  0. ]
 [ 0.   0.5]]
The mean after 1 instances is: [ 0.00907265  0.0038442 ]
The covariance matrix after 1 instances is: [[ 0.17502891 -0.13769451]
 [-0.13769451  0.44165703]]
The mean after 10 instances is: [ 0.42699157 -0.38932117]
The covariance matrix after 10 instances is: [[ 0.01707855  0.01092929]
 [ 0.01092929  0.08391719]]
The mean after 20 instances is: [ 0.39515277 -0.50837326]
The covariance matrix after 20 instances is: [[ 0.00879665  0.0032976 ]
 [ 0.0032976   0.0315392 ]]
```

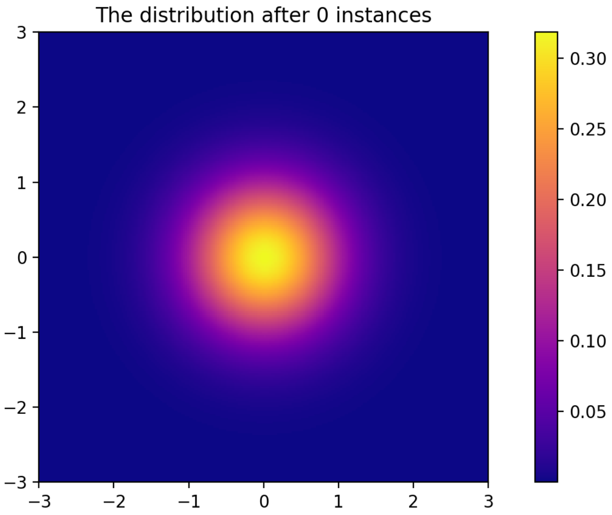


Figure 3.1: Distribution of w after 0 instances

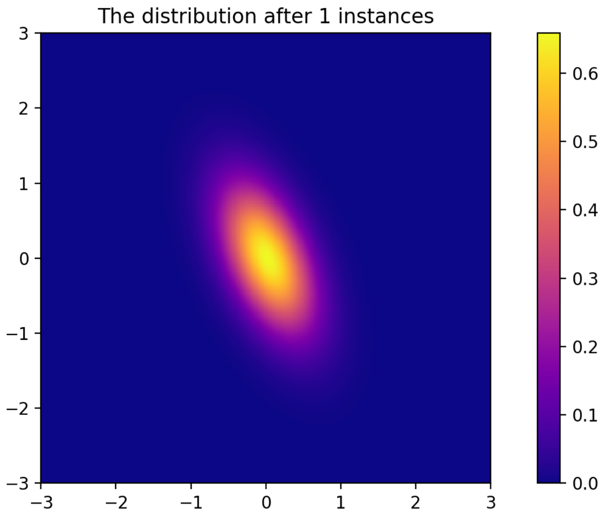


Figure 3.2: Distribution of w after 1 instances

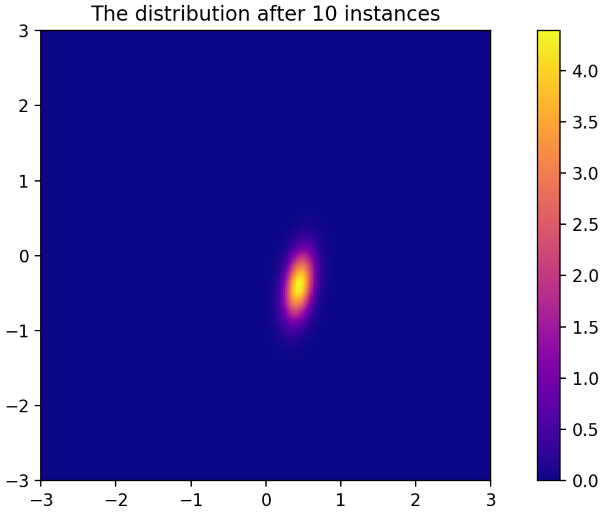


Figure 3.3: Distribution of w after 10 instances

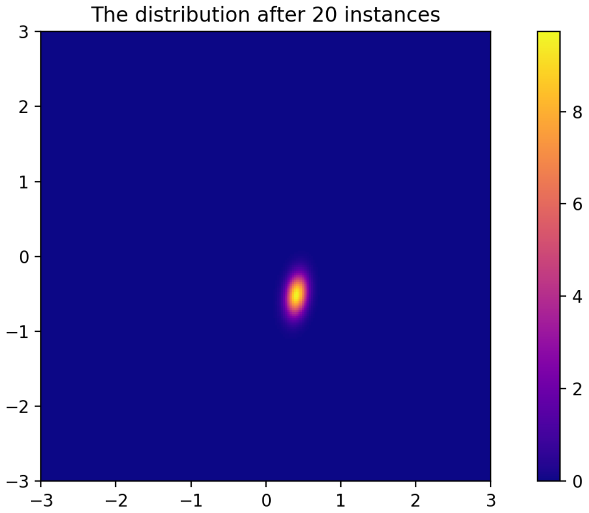


Figure 3.4: Distribution of w after 20 instances

4 Problem 4

4.1 Part A

4.1.1 I

Since we have

$$m(x) = 0, k(x, x\prime) = \exp - \frac{\|x - x\prime\|}{2\sigma^2} \tag{4.1}$$

For the joint distribution over $y(x_1), y(x_2), \cdots, y(x_n)$, we can express as:

$$y(\mathbf{x}) \sim \mathscr{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x})) \tag{4.2}$$

where

$$m(\mathbf{x}) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, K(\mathbf{x}, \mathbf{x}) = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix} \tag{4.3}$$

4.1.2 II

Five sample functions for different kernel parameters are shown in Fig 4.1, Fig 4.2 and Fig 4.3.

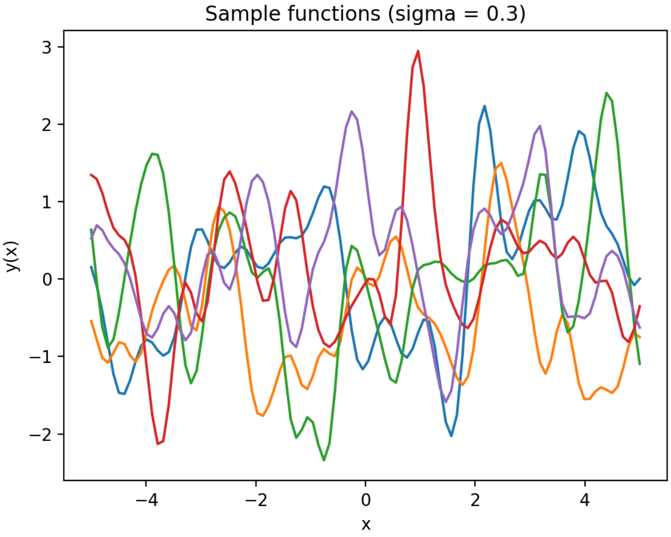


Figure 4.1: Sample functions when $\sigma = 0.3$

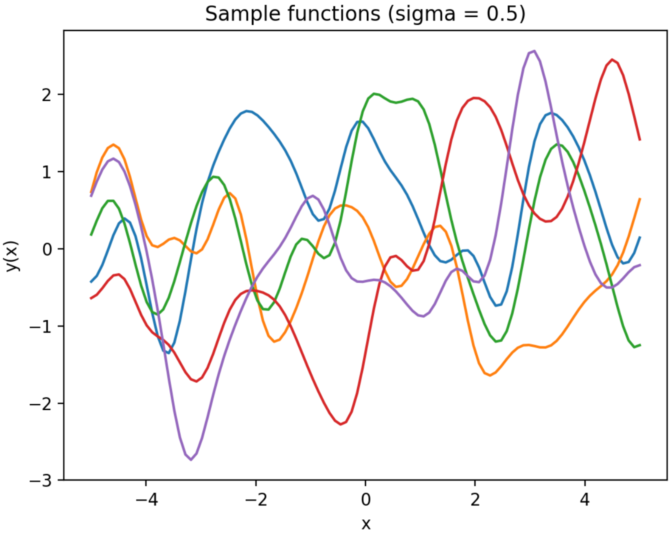


Figure 4.2: Sample functions when $\sigma = 0.5$

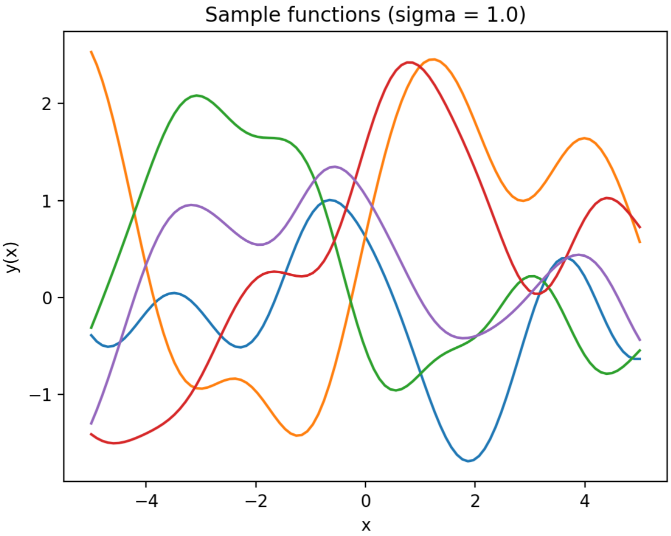


Figure 4.3: Sample functions when $\sigma = 1.0$

4.2 Part b

4.2.1 i

When we observe the input/output pairs $D = T\{(x_0^D, y_0^D), \cdots, (x_m^D, Ty_m^D)\}$, using the rules of conditioning Gaussians, we can get

$$y(x_0), y(x_1), \cdots, y(x_n)|D \sim \mathcal{N}(\mu, \sum) \tag{4.4}$$

where the mean and covariance matrix can be derived as

$$\mu = K(\mathbf{x}, \mathbf{x}^D)(K(\mathbf{x}^D, \mathbf{x}^D)^{-1}\mathbf{y}^D \tag{4.5}$$

$$\sum = K(\mathbf{x}, \mathbf{x}) - K(\mathbf{x}, \mathbf{x}^D)K(\mathbf{x}^D, \mathbf{x}^D)^{-1}K(\mathbf{x}^D, \mathbf{x}) \tag{4.6}$$

where $K(\mathbf{x}, \mathbf{x}^D) \in R^{(n+1)\times(m+1)}$, $K(\mathbf{x}^D, \mathbf{x}) \in R^{(m+1)\times(n+1)}$ and $K(\mathbf{x}^D, \mathbf{x}^D) \in R^{(m+1)\times(m+1)}$ have similar form as 4.3.

$$K(\mathbf{x}, \mathbf{x}^D) = \begin{bmatrix} k(x_0, x_0^D) & k(x_0, x_1^D) & \cdots & k(x_0, x_m^D) \\ k(x_1, x_0^D) & k(x_1, x_1^D) & \cdots & k(x_1, x_m^D) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_0^D) & k(x_n, x_1^D) & \cdots & k(x_n, x_m^D) \end{bmatrix} \tag{4.7}$$

$$K(\mathbf{x}^D, \mathbf{x}) = \begin{bmatrix} k(x_0^D, x_0) & k(x_0^D, x_1) & \cdots & k(x_0^D, x_n) \\ k(x_1^D, x_0) & k(x_1^D, x_1) & \cdots & k(x_1^D, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_m^D, x_0) & k(x_m^D, x_1) & \cdots & k(x_m^D, x_n) \end{bmatrix} \tag{4.8}$$

$$K(\mathbf{x}^D, \mathbf{x}) = \begin{bmatrix} k(x_0^D, x_0^D) & k(x_0^D, x_1^D) & \cdots & k(x_0^D, x_m^D) \\ k(x_1^D, x_0^D) & k(x_1^D, x_1^D) & \cdots & k(x_1^D, x_m^D) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_m^D, x_0^D) & k(x_m^D, x_1^D) & \cdots & k(x_m^D, x_m^D) \end{bmatrix} \tag{4.9}$$

4.2.2 II

The sample functions when given D are shown in Fig 4.4, Fig 4.5 and Fig 4.6.

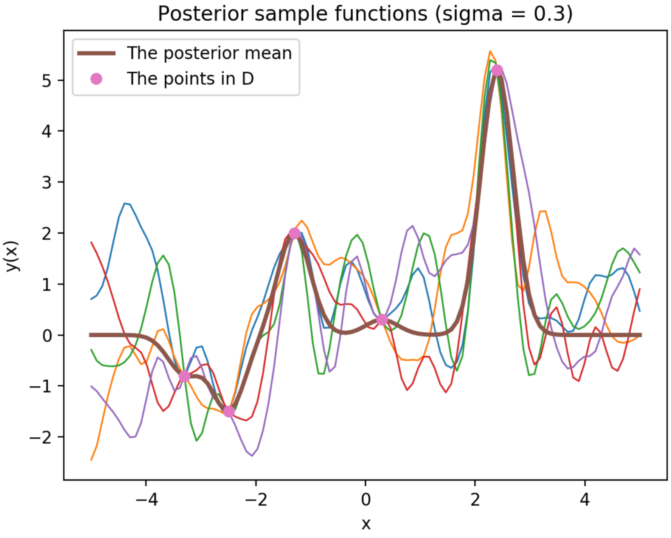


Figure 4.4: Sample functions when $\sigma = 0.3$

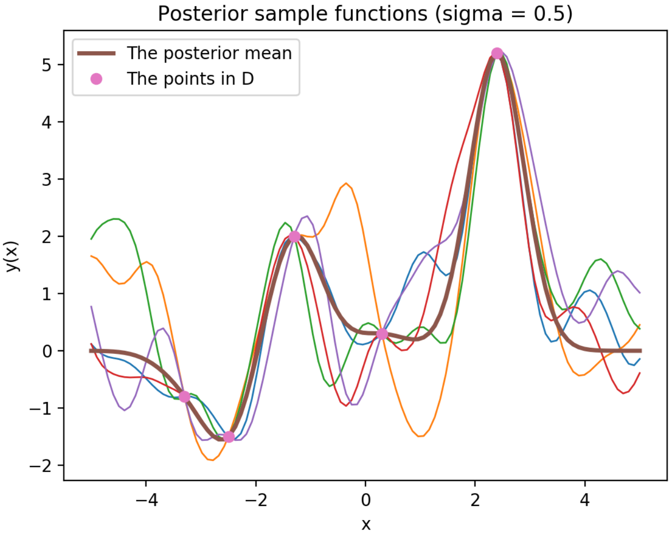


Figure 4.5: Sample functions when $\sigma = 0.5$

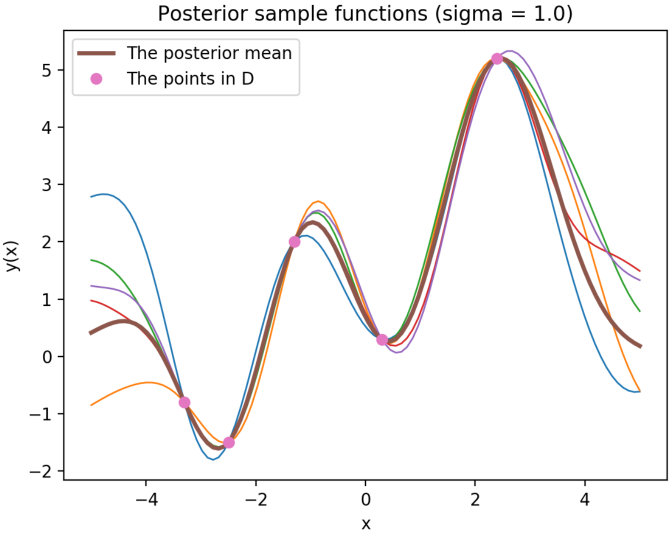


Figure 4.6: Sample functions when $\sigma = 1.0$

5 Problem 5

5.1 Part A

After training the perceptron, we can create the corresponding plot, which is shown in Fig 5.1.

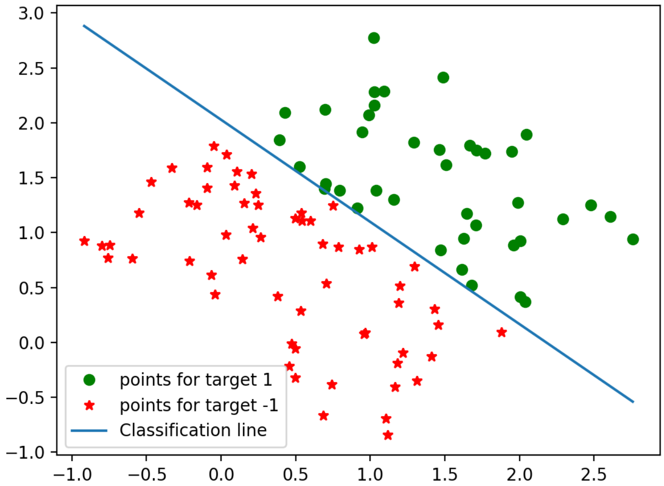
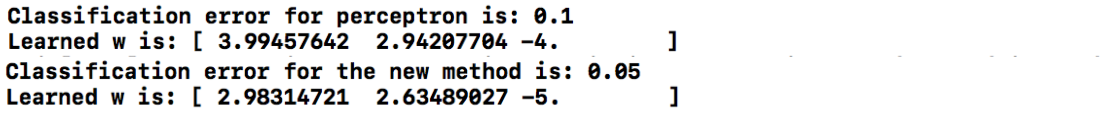


Figure 5.1: The plot of two label points and classification line

5.2 Part B

When we train the perceptron for a dataset that is not seperable, which is shown in Fig 5.2. We can use a new method, which uses the parameters of perceptron that can achieve the best classification accuracy. On each update, if the w can get the best classification accuracy, we store it. Corresponding classification error is shown as follows, and scatter plot is shown in Fig 5.3.



```
Classification error for perceptron is: 0.1
Learned w is: [ 3.99457642  2.94207704 -4.        ]
Classification error for the new method is: 0.05
Learned w is: [ 2.98314721  2.63489027 -5.        ]
```

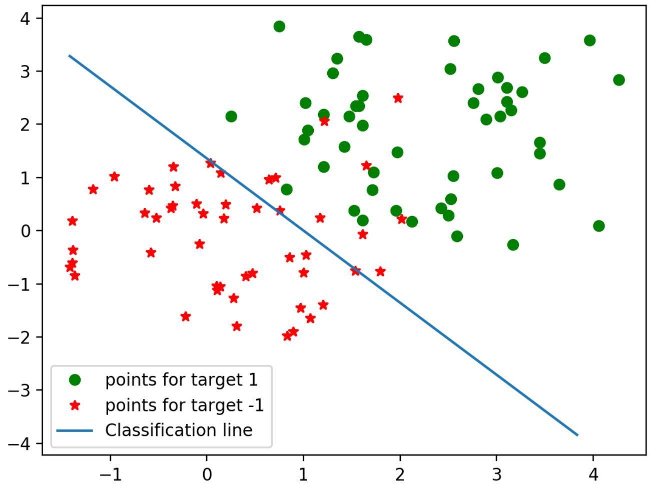


Figure 5.2: The plot of two label points and classification line (not seperable)

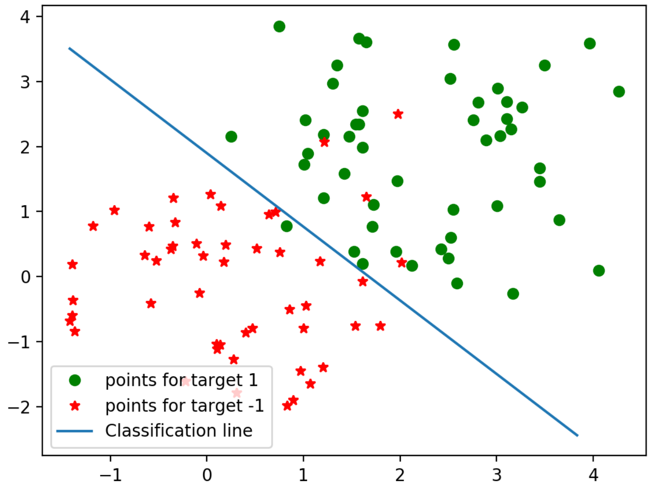
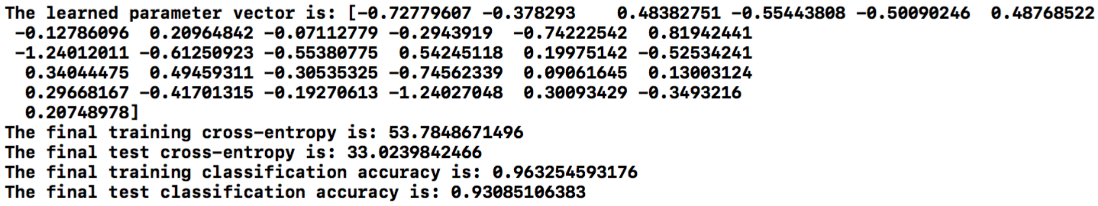


Figure 5.3: The plot of two label points and classification line for the new method

6 Problem 6

In this problem, we use the logistic regression for the breast cancer wisconsin dataset. The training and test calssification accuracy vs. the number of iterations of SGD is shown in Fig 6.1. The average training and test error vs. the number of iterations of SGD is shown in Fig 6.2. The learned 31-dimensional parameter vector w (includes bias term) and final training and test cross-entropy and classification accuracy are:



```
The learned parameter vector is: [-0.72779607 -0.378293    0.48382751 -0.55443808 -0.50090246  0.48768522
 -0.12786096  0.20964842 -0.07112779 -0.2943919  -0.74222542  0.81942441
 -1.24012011 -0.61250923 -0.55380775  0.54245118  0.19975142 -0.52534241
  0.34044475  0.49459311 -0.30535325 -0.74562339  0.09061645  0.13003124
  0.29668167 -0.41701315 -0.19270613 -1.24027048  0.30093429 -0.3493216
  0.20748978]
The final training cross-entropy is: 53.7848671496
The final test cross-entropy is: 33.0239842466
The final training classification accuracy is: 0.963254593176
The final test classification accuracy is: 0.93085106383
```

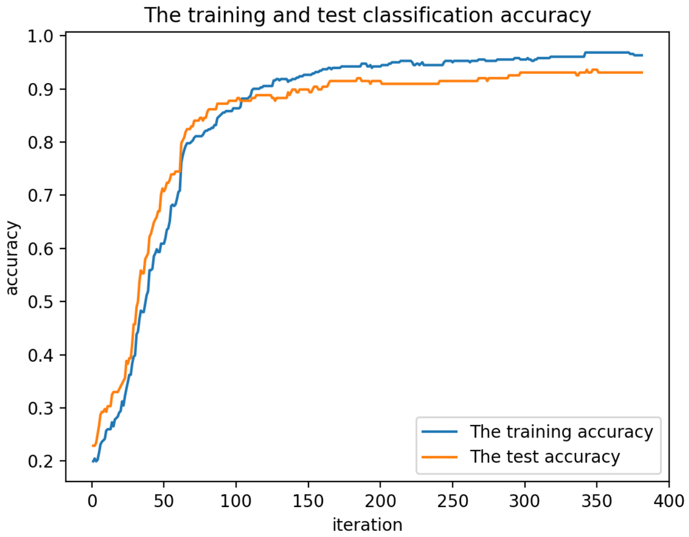


Figure 6.1: The training and test accuracy

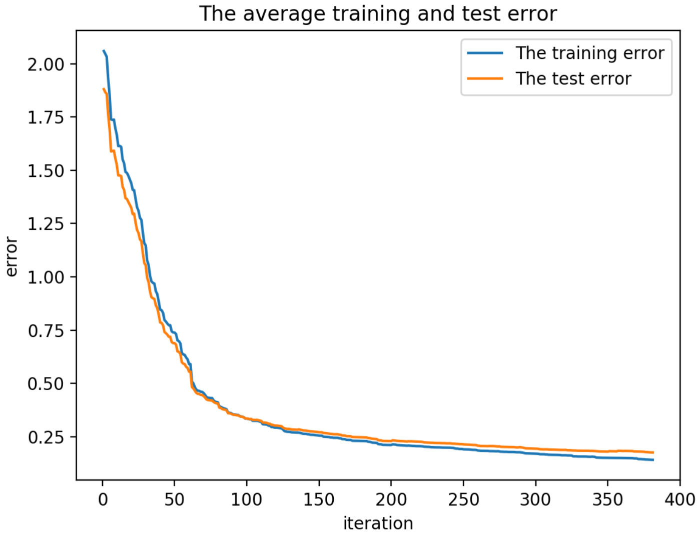


Figure 6.2: The average training and test error