Problem 1

(a)

For class means μ_1, μ_2, we can predict $\hat{y}(\underline{x})$ by taking $\arg\min_k ||\underline{x} - \mu_k||_2^2, k = 1, 2$

Let $\tilde{g}_k(\underline{x}) = -||\underline{x} - \mu_k||^2$, simplify it we can get that $\tilde{g}_k(\underline{x}) = -(\underline{x}^T\underline{x} - 2\mu_k^T\underline{x} + \mu_k^T\mu_k)$, let $g_k(\underline{x}) = 2\mu_k^T\underline{x} - \mu_k^T\mu_k$

Thus, we get the $g(\underline{x})$ of two-class nearest means classifier: $g_k(\underline{x}) = 2\mu_k^T\underline{x} - \mu_k^T\mu_k, k = 1, 2$

(b)

$$
\begin{aligned}
g_{Dk}(x) &= 2\mu_k^T\underline{x} - \mu_k^T\mu_k \\
&= -\frac{1}{N_k^2}(\sum_{j=1}^{N_k}\underline{x}_i^{(k)^T})(\sum_{i=1}^{N_k}\underline{x}_j^{(k)}) + \frac{2}{N_k}\sum_{i=1}^{N_k}\underline{x}_i^{(k)^T}x \qquad (1) \\
k &= 1, 2
\end{aligned}
$$

Dual representation:

$g_{D1}(\underline{x}) = -\frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\underline{x}_i^{(1)^T}\underline{x}_j^{(1)} + \frac{2}{N_1}\sum_{i=1}^{N_1}\underline{x}_i^{(1)^T}\underline{x}$

$g_{D2}(\underline{x}) = -\frac{1}{N_2^2}\sum_{i=1}^{N_2}\sum_{j=1}^{N_2}\underline{x}_i^{(2)^T}\underline{x}_j^{(2)} + \frac{2}{N_2}\sum_{i=1}^{N_2}\underline{x}_i^{(2)^T}x$

(c)

RBF kernel is defined as $K(x, x^{'}) = \exp\left(-\gamma||x - x^{'}||_2^2\right)$

Then, we can get

$$
\begin{aligned}
g_{D1}(\underline{x}) &= -\frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}K(\underline{x}_i^{(1)}, \underline{x}_j^{(1)}) + \frac{2}{N_1}\sum_{i=1}^{N_1}K(\underline{x}_i^{(1)}, \underline{x}) \\
&= -\frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\exp\left(-\gamma||\underline{x}_i^{(1)} - \underline{x}_j^{(1)}||_2^2\right) + \frac{2}{N_1}\sum_{i=1}^{N_1}\exp\left(-\gamma||\underline{x}_i^{(1)} - \underline{x}||_2^2\right) \qquad (2)
\end{aligned}
$$

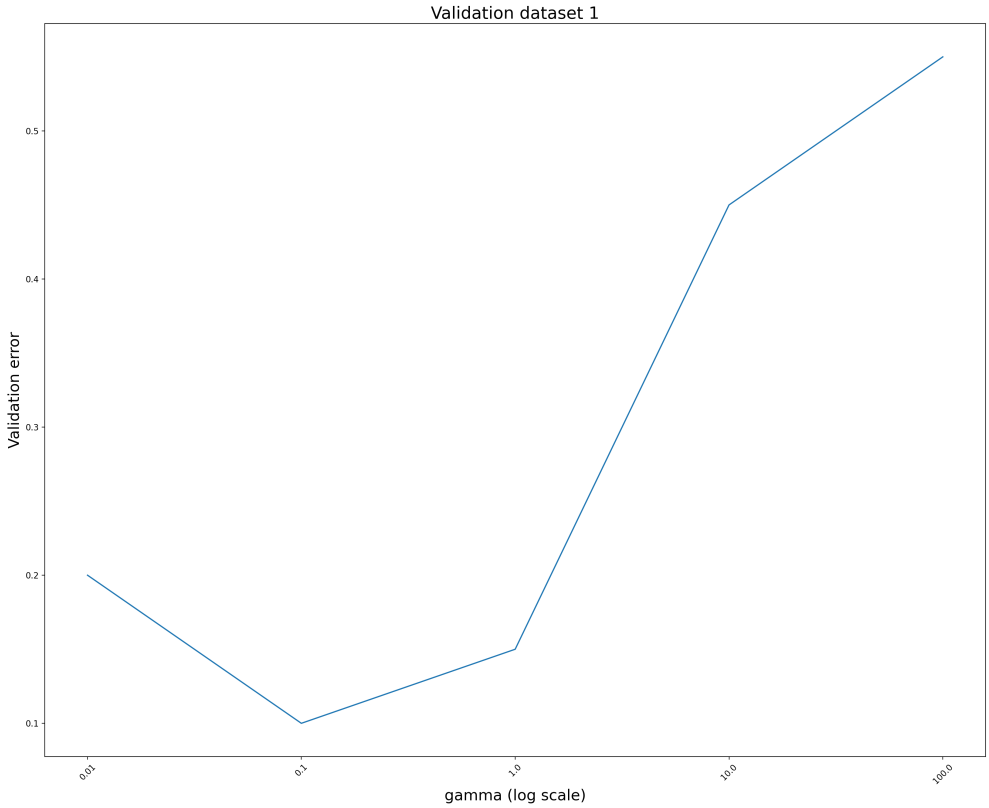
$$
\begin{aligned}
g_{D2}(\underline{x}) &= -\frac{1}{N_2^2}\sum_{i=1}^{N_2}\sum_{j=1}^{N_2}K(\underline{x}_i^{(2)}, \underline{x}_j^{(2)}) + \frac{2}{N_2}\sum_{i=1}^{N_2}K(\underline{x}_i^{(2)}, x) \\
&= -\frac{1}{N_2^2}\sum_{i=1}^{N_2}\sum_{j=1}^{N_2}\exp\left(-\gamma||\underline{x}_i^{(2)} - \underline{x}_j^{(2)}||_2^2\right) + \frac{2}{N_2}\sum_{i=1}^{N_2}\exp\left(-\gamma||\underline{x}_i^{(2)} - x||_2^2\right) \qquad (3)
\end{aligned}
$$

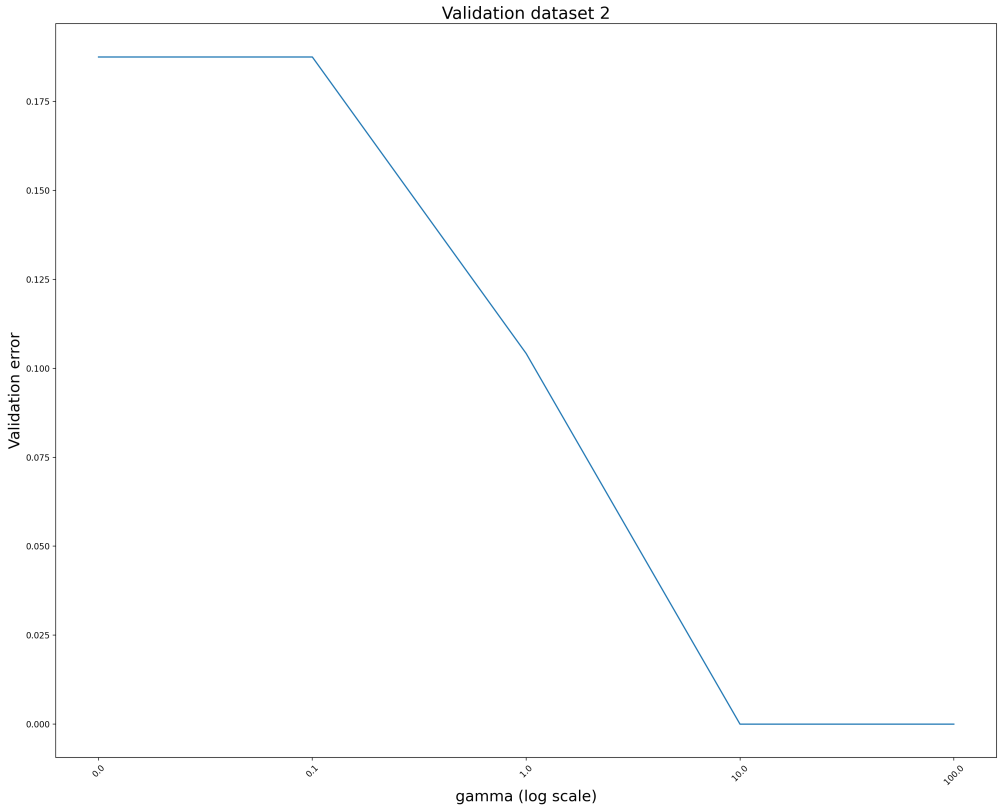
(d)

For dataset 1, the best γ for validation set is 0.1.

For dataset 2, the best γ for validation set is 10 or 100.

(e)


Validation dataset 1
Validation error
0.5
0.4
0.3
0.2
0.1
0.01
0.1
1.0
10.0
100.0
gamma (log scale)


Validation dataset 2
Validation error
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.0
0.1
1.0
10.0
100.0
gamma (log scale)

(f)

```
The test error of linear kernel on the dataset1 is 0.4300
The test error of linear kernel on the dataset2 is 0.2438
The test error of rbf kernel on the dataset1 is 0.2400
The test error of rbf kernel on the dataset2 is 0.0000
```

The test error of linear kernel on the dataset1 is 0.4300.

The test error of linear kernel on the dataset2 is 0.2438.

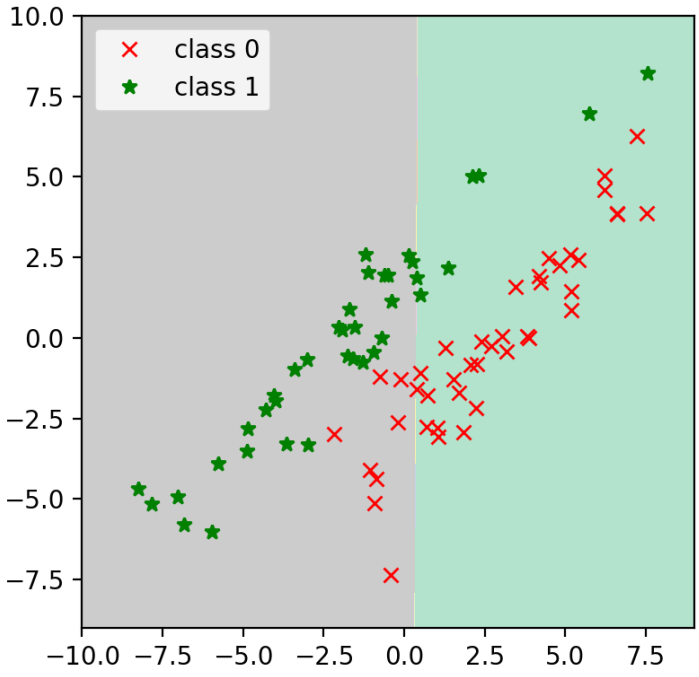
The test error of rbf kernel on the dataset1 is 0.2400.

The test error of rbf kernel on the dataset2 is 0.0000.

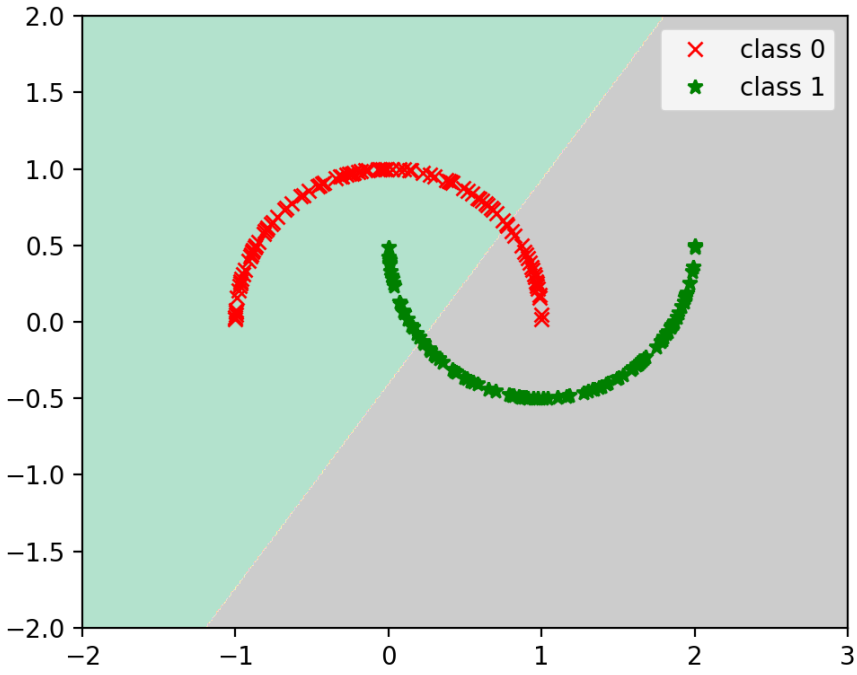
From the results, we can find that the error of rbf kernel is lower than the linear kernel method.

(g)

For dataset 1

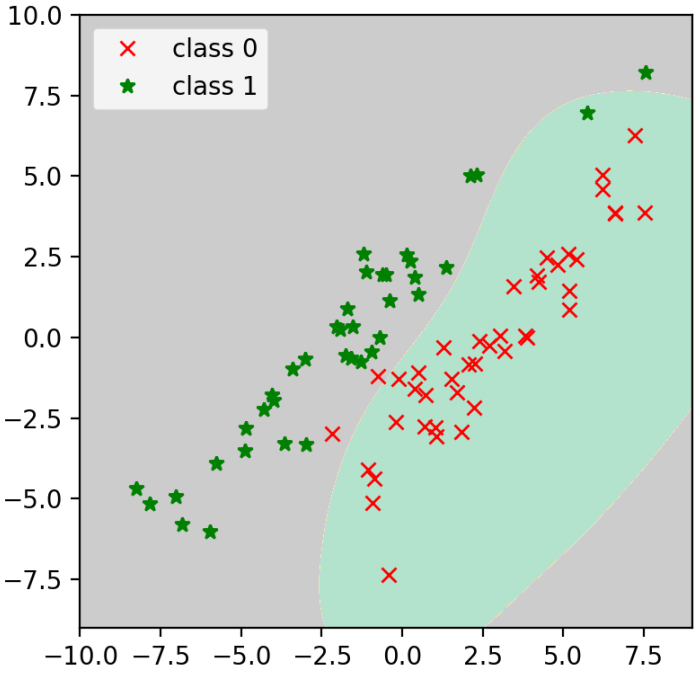


For dataset 2

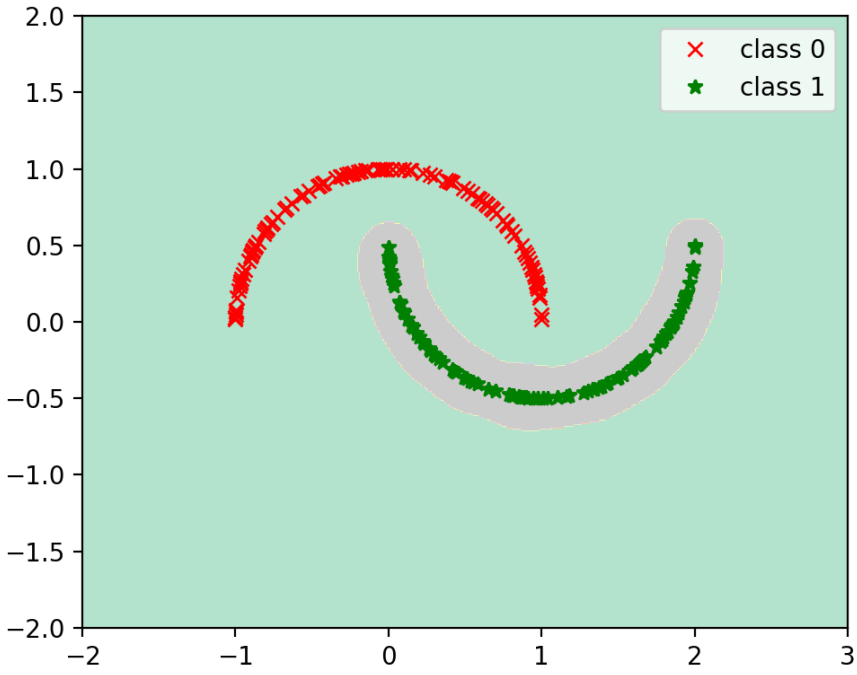


(h)

For dataset 1

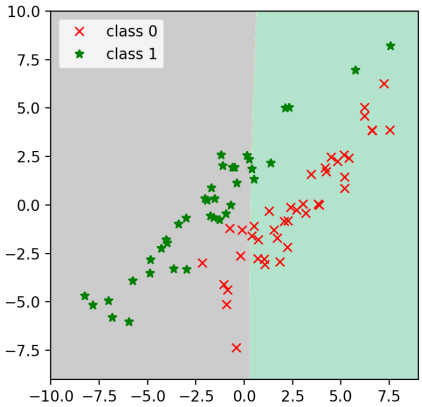


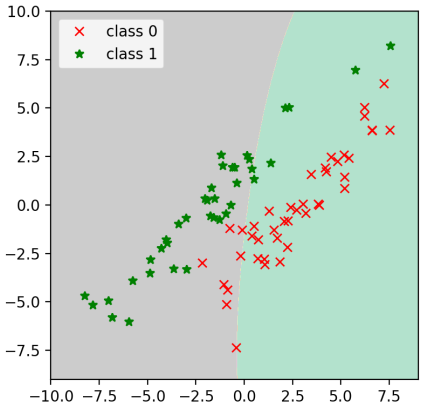
For dataset 2

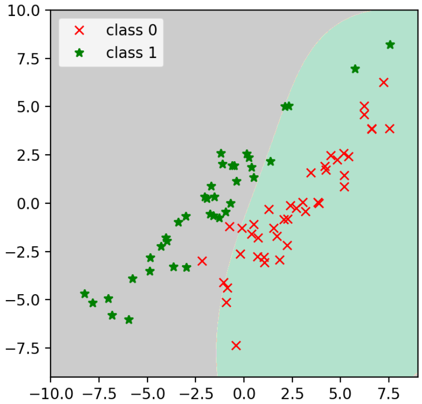


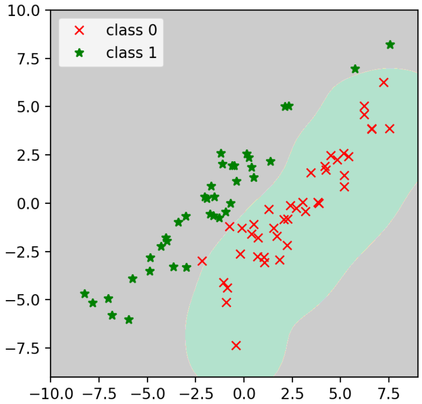
(i)

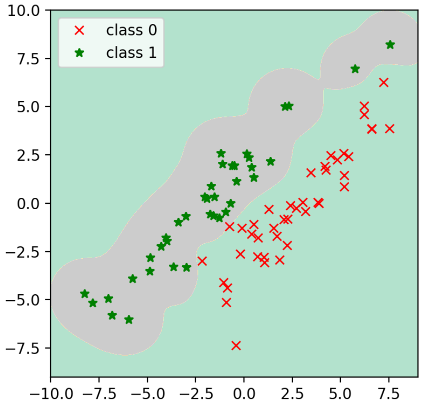
For dataset 1

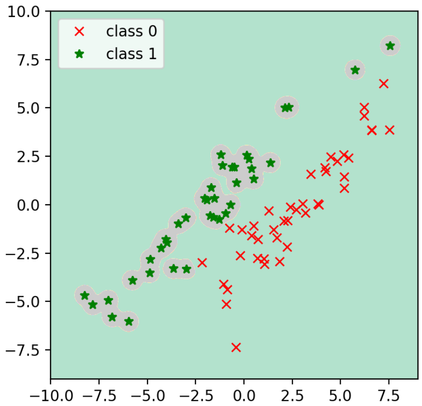




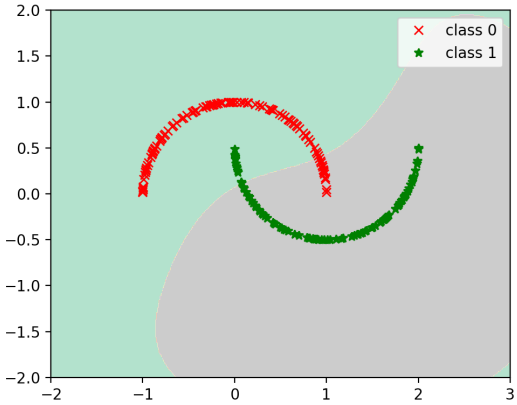


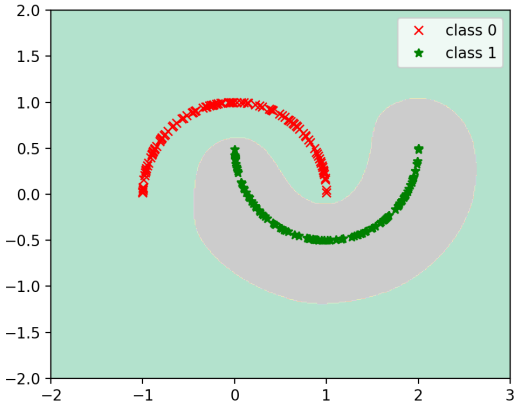


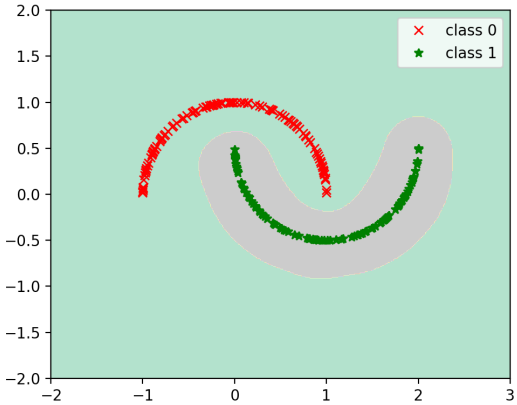


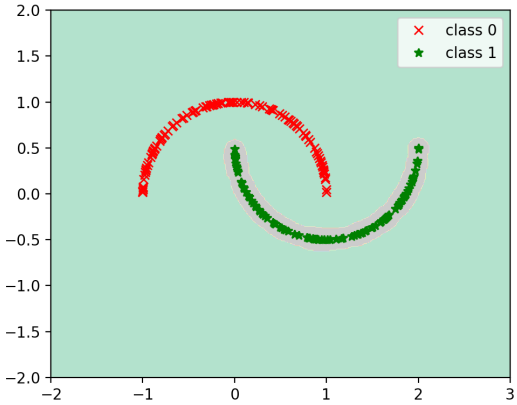


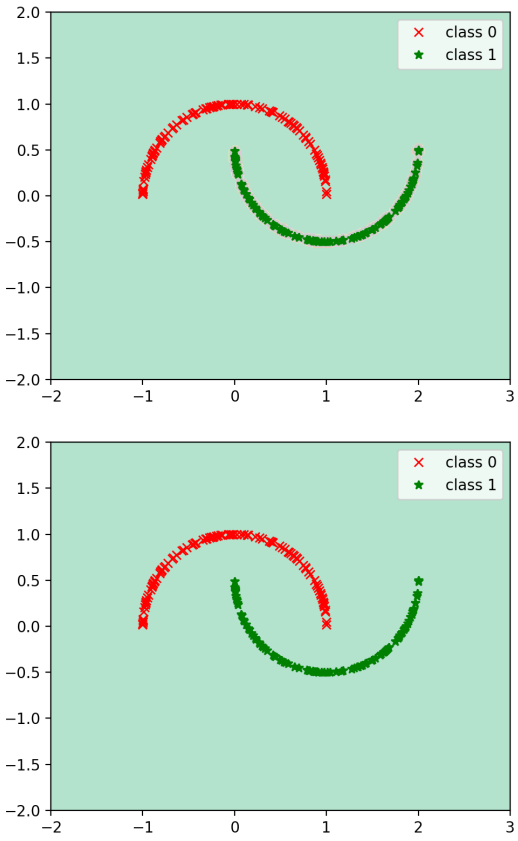
For dataset 2


class 0
class 1
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2
3


class 0
class 1
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2
3


class 0
class 1
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2
3


class 0
class 1
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2
3



From the figure above, we can find that when we increase the γ, the decision boundary will become more tortuous, and the data can be separated well due to the complicated decision boundary. But it also increases the risk of overfitting problem.

Problem 2

(a)

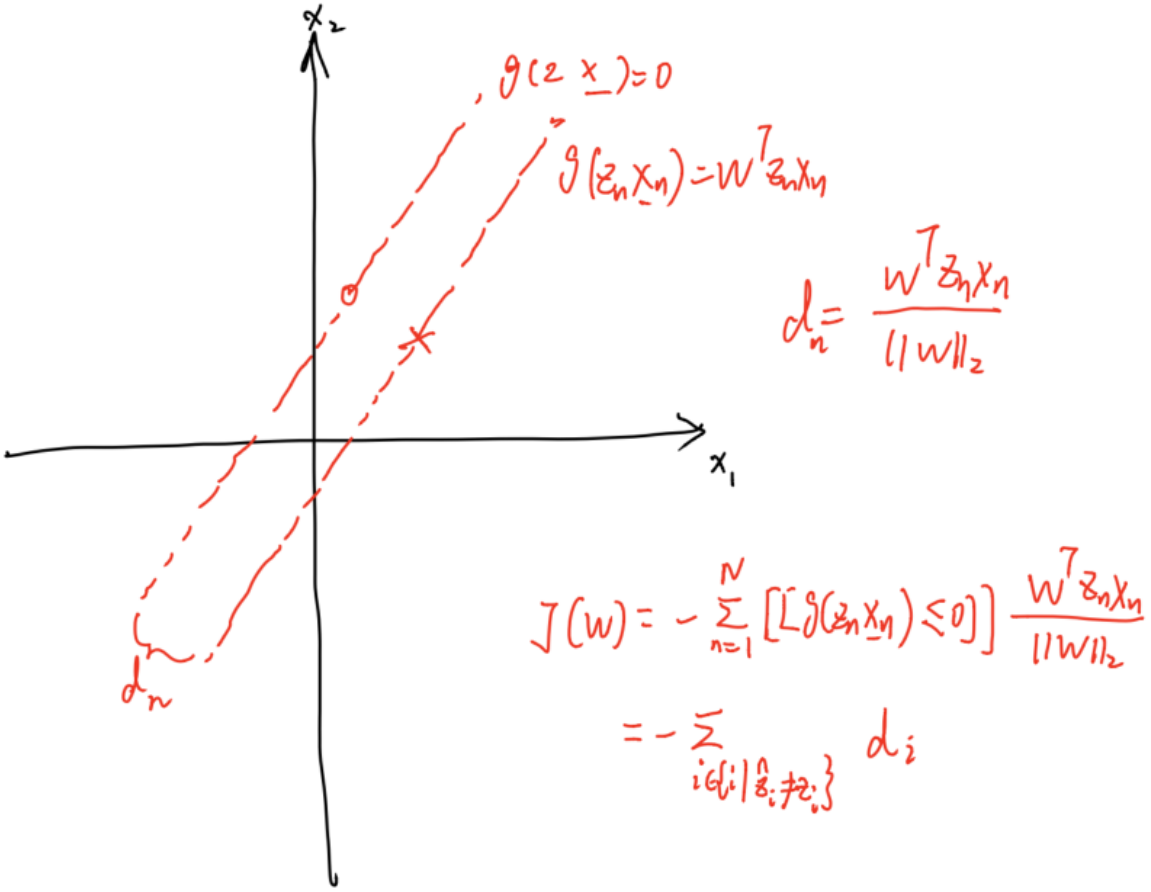
The objective is

$$J(\underline{w}) = -\sum_{n=1}^{N}[[\underline{w}^T z_n \underline{x}_n \leq 0]] \frac{\underline{w}^T z_n \underline{x}_n}{||\underline{w}||_2} \tag{4}$$

We know that the algorithm classifies the data point correctly if $\underline{w}^T z_n \underline{x}_n \geq 0$, therefore in this objective, it only considers the errors taken by the misclassified data points. Thus, it can be simplified as

$$J(\underline{w}) = -\sum_{i \in \{i|\hat{z}_i \neq z_i\}} \frac{\underline{w}^T z_i \underline{x}_i}{||\underline{w}||_2} \tag{5}$$

where $\{i|\hat{z}_i \neq z_i\}$ denotes the set which contains all the misclassified data points.



In the figure, we can find that the d_i is the distance from the misclassified prediction to the decision boundary, to optimize the objective function we can make the gap become small and increase our model's accuracy.

(b)

Batch GD:

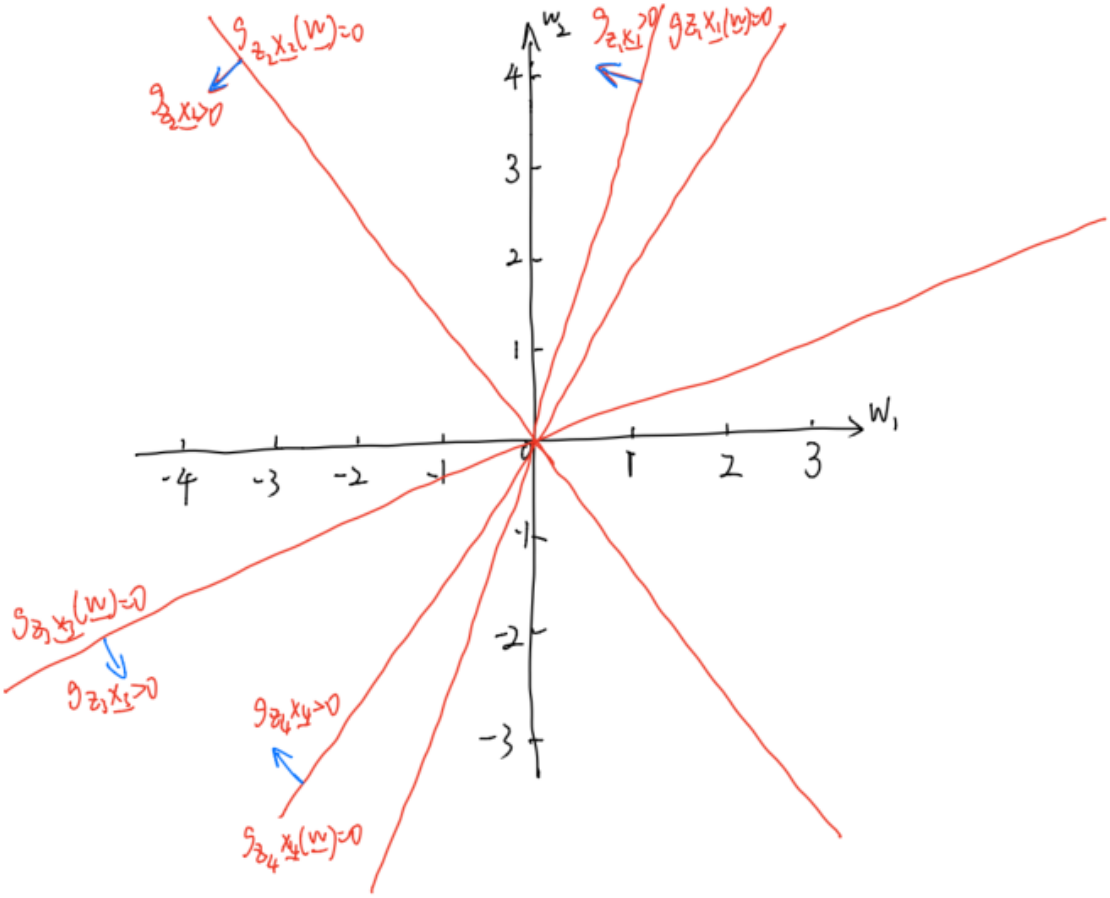
- initialize $\underline{w}(0) = 0$
- Loop:
  - compute $g(z_n \underline{x}_n)$ on the whole dataset, and keep the data points in set A which satisfy $g(z_n \underline{x}_n) \le 0$
  - compute the gradient updates using set A: $\nabla_{\underline{w}} J = -\frac{1}{|A|} \sum_{x_n \in A} \frac{||\underline{w}||_2^2 I - \underline{w}\underline{w}^T}{||\underline{w}||_2^3} z_n \underline{x}_n$
  - update the weight : $\underline{w}(i+1) \leftarrow \underline{w}(i) - \eta \nabla_{\underline{w}} J$
  - if convergence, end loop

Stochastic GD:

- initialize $\underline{w}(i) = 0$
- Loop:
  - for x_n in dataset:
    - if $g(z_n \underline{x}_n) \le 0$:
      - compute the gradient updates using x_n: $\nabla_{\underline{w}} J = -\frac{||\underline{w}||_2^2 I - \underline{w}\underline{w}^T}{||\underline{w}||_2^3} z_n \underline{x}_n$
      - update the weight : $\underline{w}(i+1) \leftarrow \underline{w}(i) - \eta \nabla_{\underline{w}} J$
- if convergence, end loop

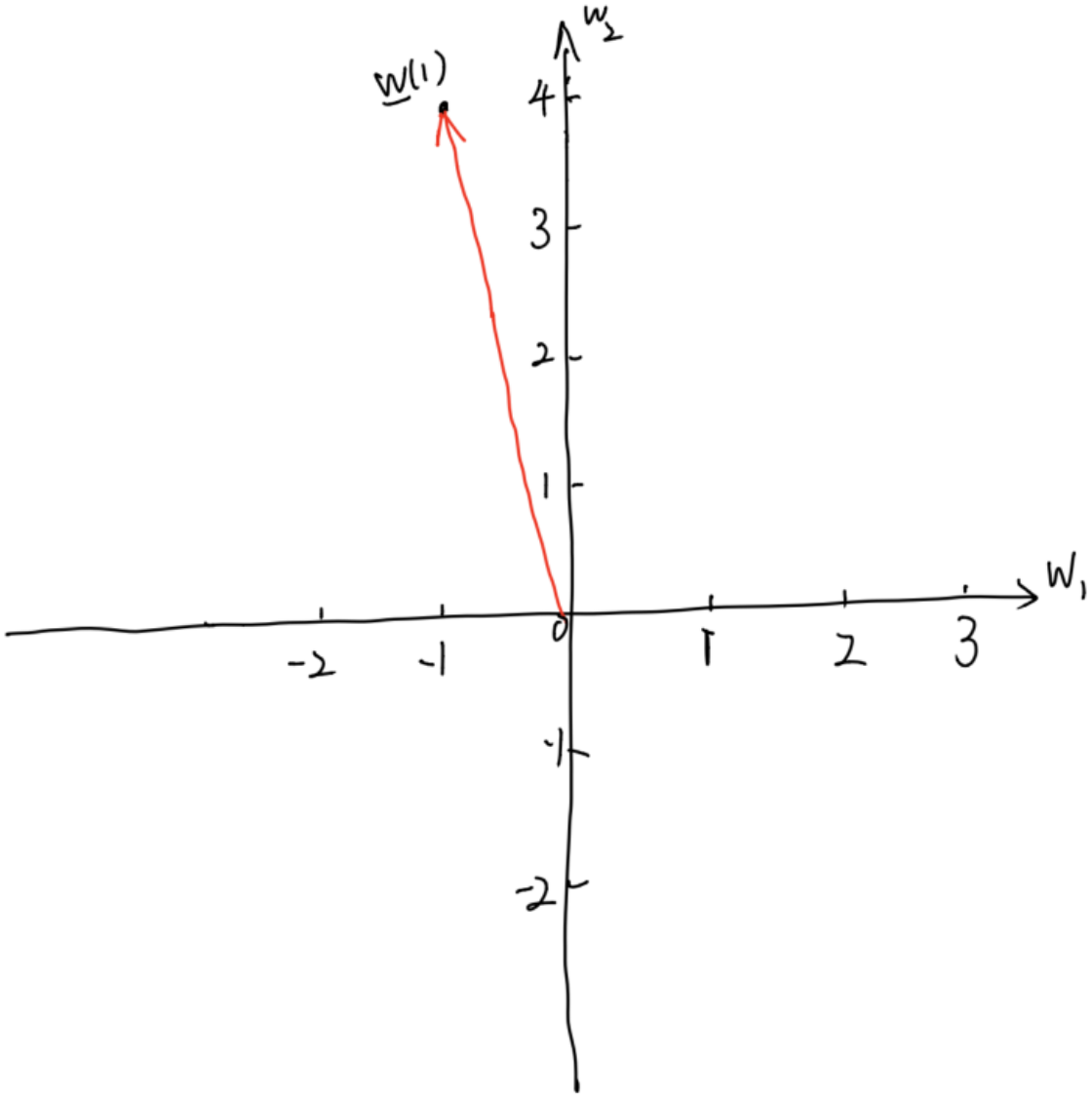
(c)

(i)



(ii)

The weight vector $\underline{w}(1)$

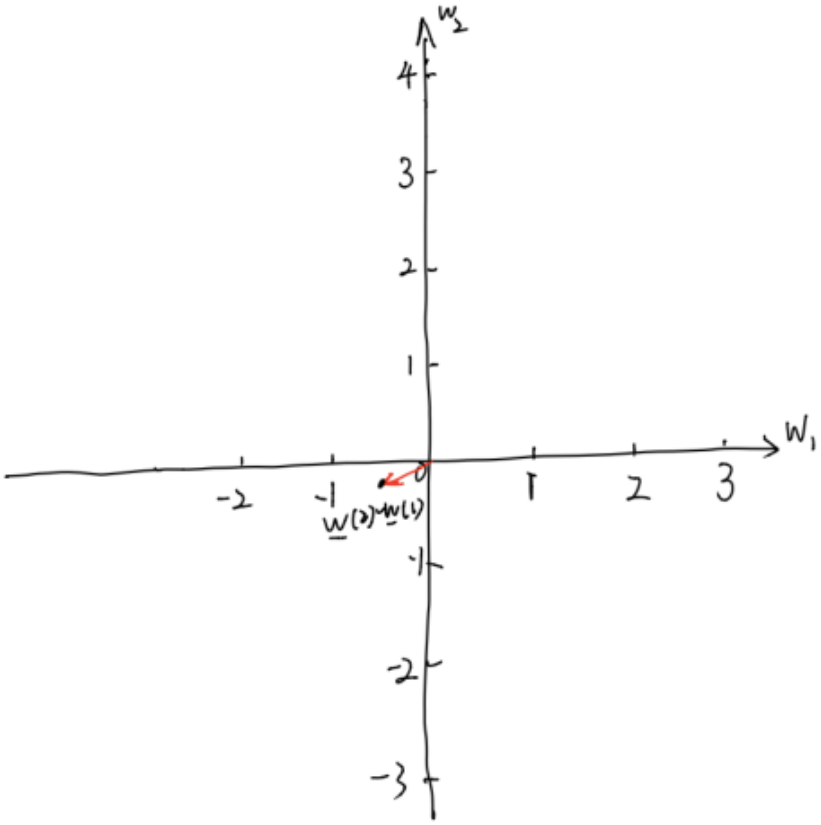


(iii)

In iteration 1, using point $\underline{x}_2$ to compute the weight update as follows:

$$\begin{aligned}\Delta\underline{w} &= -\eta(i)\nabla_{\underline{w}}J \\ &= (-\frac{1}{i+1})(-\frac{||\underline{w}||_2^2 I - \underline{w}\underline{w}^T}{||\underline{w}||_2^3} z_n \underline{x}_n) \\ &= -\frac{1}{2}\begin{bmatrix}\frac{56\sqrt{17}}{289} \\ \frac{14\sqrt{17}}{289}\end{bmatrix} \\ &= [-0.4, -0.1]^T\end{aligned} \tag{6}$$

Thus, we can get that weight update $\underline{w}(2) - \underline{w}(1) = [-0.4, -0.1]^T$.

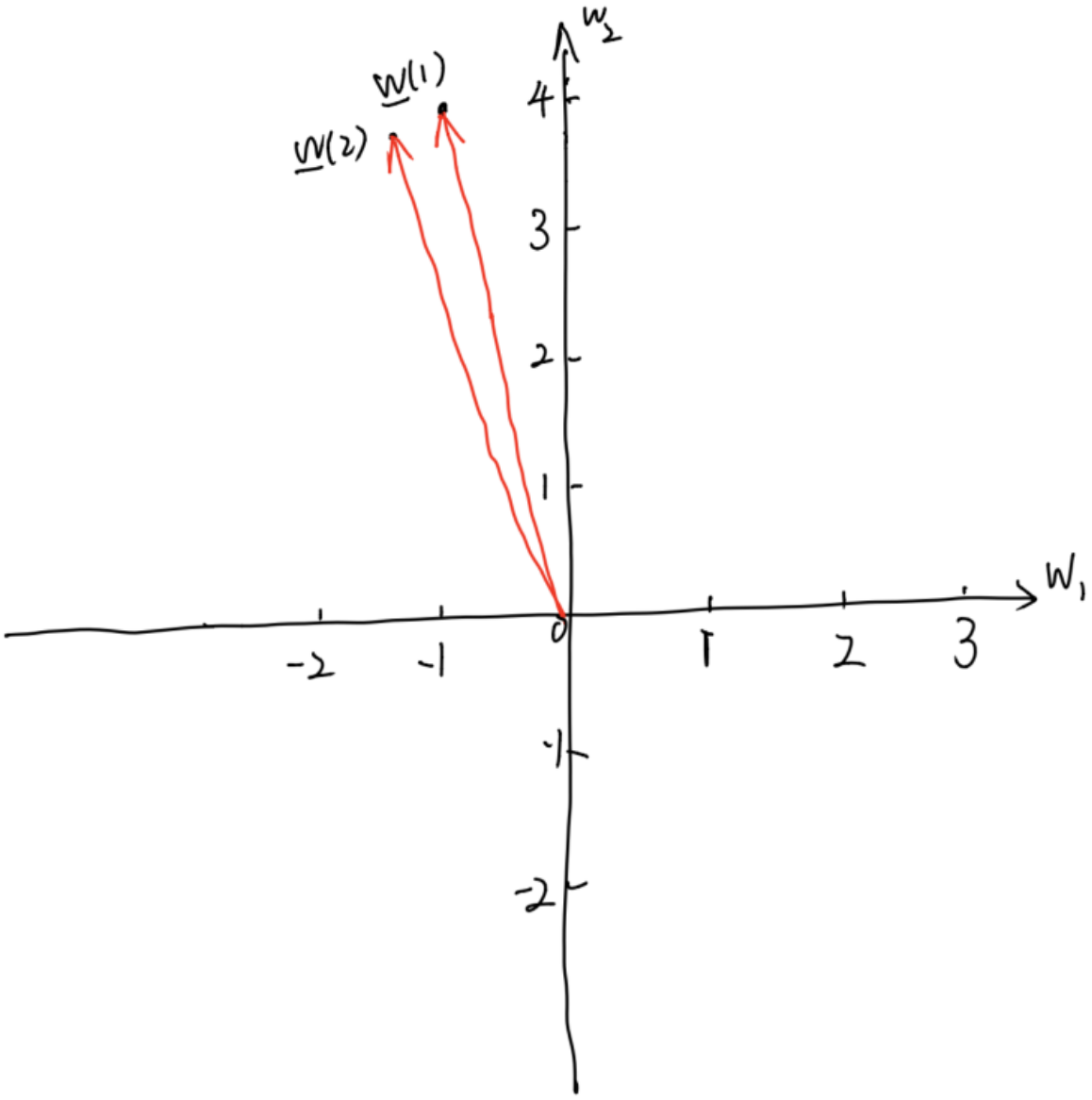


(iv)

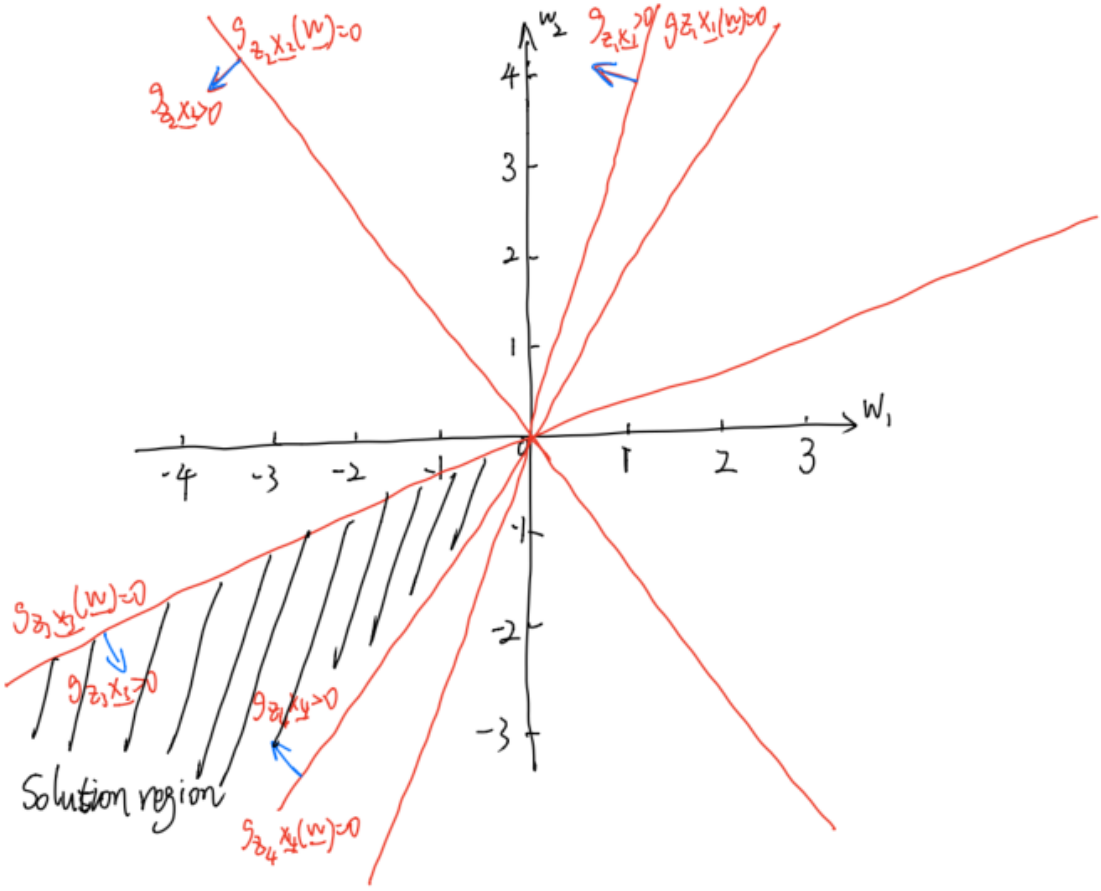
From the derivation above, we can compute the updated weight as follows:

$$\begin{aligned} \underline{w}(2) \leftarrow & \underline{w}(1) - \frac{1}{2}\nabla_{\underline{w}} J \\ \underline{w}(2) \leftarrow & \underline{w}(1) - \frac{1}{2}\begin{bmatrix} \frac{56\sqrt{17}}{289} \\ \frac{14\sqrt{17}}{289} \end{bmatrix} \end{aligned} \tag{7}$$

Therefore, we can get that $\underline{w}(2) = [-1.4, 3.9]^T$.



(v)



There exists solution region as the figure above show.

(d)

The criterion function is convex.

Assume we take one point x_n into consideration, first we compute the first-order derivation of $J = -\frac{w^T z_n x_n}{||w||_2}$

$$J^{'} = -\frac{I}{||w||_2} + \frac{ww^T}{||w||_2^3} \tag{8}$$

Then, we derive the second-order derivation of it.

$$\begin{aligned} J^{''} &= \frac{w}{||w||_2^3} + \frac{2w||w||_2^3 - 3w||w||_2 w^T w}{||w||_2^6} \\ &= \frac{3w}{||w||_2^3} - \frac{3ww^T w}{||w||_2^5} \\ &\geq 0 \end{aligned} \tag{9}$$

Because the second-order derivation of the criterion function is greater than 0, the criterion function is convex function.

Problem 3

(a)

The objective is listed as follow:

$$J_{A.1}(\underline{w}^{'}) = ||\underline{w}^{'}\underline{u} - y||_2^2 \tag{10}$$

where
$\underline{u} = [1, \underline{x_1}, \underline{x_1}^2, \underline{x_1}^3, \underline{x_2}, \underline{x_2}^2, \underline{x_2}^3, \underline{x_3}, \underline{x_3}^2, \underline{x_3}^3, \underline{x_1x_2}, \underline{x_1x_3}, \underline{x_2x_3}, \underline{x_1x_2}^2, \underline{x_1x_3}^2, \underline{x_2x_1}^2, \underline{x_2x_3}^2, \underline{x_3x_1}^2, \underline{x_3x_2}^2, \underline{x_1x_2x_3}]$
, y is the true value we need to fit.

Because $\underline{u}$ has 20 features, $\underline{w}^{'}$ also has 20 components.

(b)

There are 20 d.o.f during learning using Approach A.1, because the model is established to fit 20 augmented features.

(c)

The optimal solution is $\underline{\hat{w}}^{'} = (\underline{u}^T\underline{u})^{-1}\underline{u}^T y$

where
$\underline{u} = [1, \underline{x_1}, \underline{x_1}^2, \underline{x_1}^3, \underline{x_2}, \underline{x_2}^2, \underline{x_2}^3, \underline{x_3}, \underline{x_3}^2, \underline{x_3}^3, \underline{x_1x_2}, \underline{x_1x_3}, \underline{x_2x_3}, \underline{x_1x_2}^2, \underline{x_1x_3}^2, \underline{x_2x_1}^2, \underline{x_2x_3}^2, \underline{x_3x_1}^2, \underline{x_3x_2}^2, \underline{x_1x_2x_3}]$
, y is the true value we need to fit.

(d)

The objective is listed as follow:

$$J_{A.2}(\underline{w}, \underline{w}^{'}) = ||\hat{y}(\underline{x}) - y||_2^2 \tag{11}$$

where $\hat{y}(\underline{x}) = w_1^{'}(\hat{f}(\underline{x})) + w_2^{'}(\hat{f}(\underline{x}))^2 + w_3^{'}(\hat{f}(\underline{x}))^3$, $\hat{f}(\underline{x}) = \underline{w}^T\underline{x}$, and y is the true value.

(e)

There are 21 d.o.f during learning using Approach A.2, because $\hat{y}(\underline{x})$ is defined with 20 degrees of freedom, and $\hat{f}(\underline{x})$ is also defined with a degree of freedom.

(f)

In Approach A.2, $\hat{y}(\underline{x})$ is a nonlinear function of $\underline{x}$, $\hat{y}(\underline{x})$ is a nonlinear function of $\underline{w}$, and $\hat{y}(\underline{x})$ is a linear function of $\underline{w}^{'}$

(g)

The criterion function of (d) is a continuous, differential function of $\underline{w}$, and is a continuous, differential function of $\underline{w}^{'}$.

(h)

For $\underline{w}$ updating,

$$\begin{aligned} \frac{\partial \hat{y}(\underline{x})}{\partial \underline{w}} &= \frac{\partial \hat{y}(\underline{x})}{\partial \hat{f}(\underline{x})}\frac{\partial \hat{f}(\underline{x})}{\partial \underline{w}} \\ &= [\underline{w}_1^{'} + 2\underline{w}_2^{'}\hat{f}(\underline{x}) + 3\underline{w}_3^{'}\hat{f}(\underline{x})^2]x^T \end{aligned} \tag{12}$$

$$\underline{w} \leftarrow \underline{w} - \mu(i)\frac{\partial \hat{y}(\underline{x})}{\partial \underline{w}} \tag{13}$$

For $\underline{w}^{'}$ updating,

$$\begin{aligned} \frac{\partial \hat{y}(\underline{x})}{\partial \underline{w}^{'}} &= [\hat{f}(\underline{x}), \hat{f}(\underline{x})^2, \hat{f}(\underline{x})^3] \\ \underline{w}^{'} &\leftarrow \underline{w}^{'} - \rho(i)\frac{\partial \hat{y}(\underline{x})}{\partial \underline{w}^{'}} \end{aligned} \tag{14}$$

(i)

The objective for Approach A.1 that also includes l_2 regularization of all the weights is listed as follow:

$$J_{A.1'}(\underline{w}^{'}) = ||\underline{w}^{'}\underline{u} - y||_2^2 + \lambda||\underline{w}^{'}||_2^2 \tag{15}$$

(j)

The objective for Approach A.1 that also includes l_2 regularization of all the weights is listed as follow:

$$J_{A.2'}(\underline{w}, \underline{w}^{'}) = ||\hat{y}(\underline{x}) - y||_2^2 + \lambda||\underline{w}||_2^2 + \lambda^{'}||\underline{w}^{'}||_2^2 \tag{16}$$

where $\hat{y}(\underline{x}) = w_1^{'}(\hat{f}(\underline{x})) + w_2^{'}(\hat{f}(\underline{x}))^2 + w_3^{'}(\hat{f}(\underline{x}))^3$, $\hat{f}(\underline{x}) = \underline{w}^T\underline{x}$, and y is the true value.

(k)

Approach A.2 is likely to give lower error on unknowns if there is plenty of data. Because the model is more complex than the approach A.1, it will have more powerful ability to capture the data's information when the data amount is large enough.

(l)

Approach is likely to give lower error on unknowns if there is a very limited amount of data. The model is easy than the approach A.2, it will not be easy to overfit to the limited dataset which makes it may have lower errors on the unknowns.