Statistical Pattern Recognition Hw6

Name: Chang-Hong Chen

UID: 117397857

Problem 1.

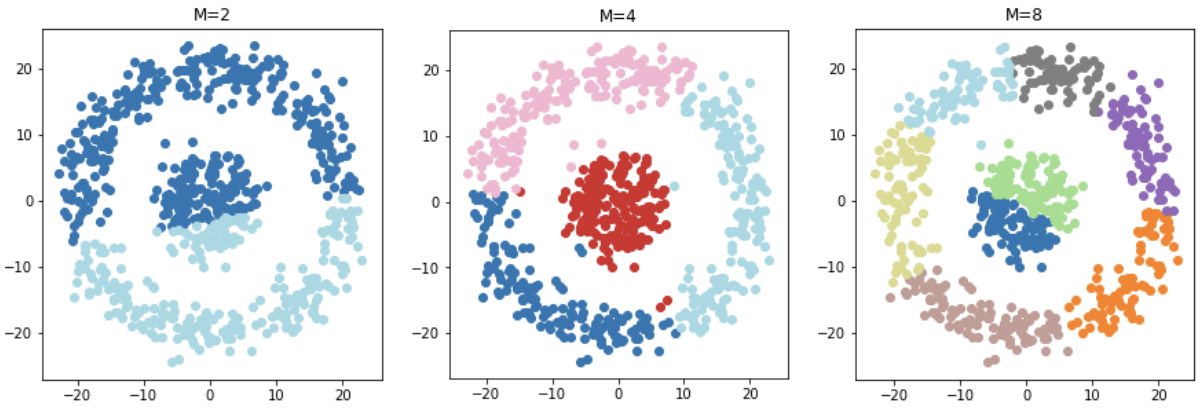
Answer:

(a)

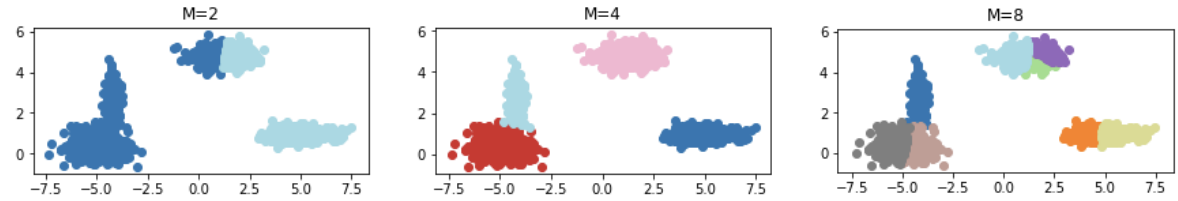
K-means

The K-means algorithm runs 3 iteration for every M. And the initial center is random points in the data.

Data 1



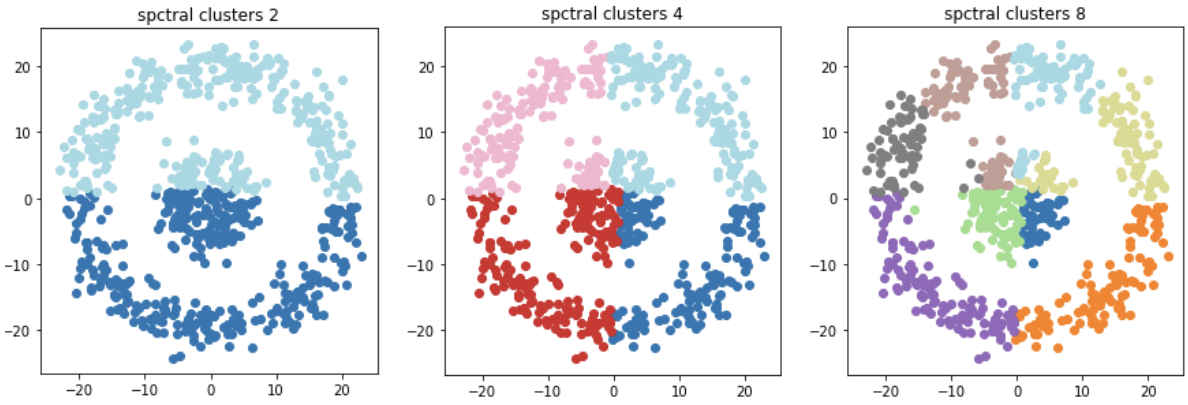
Data 2



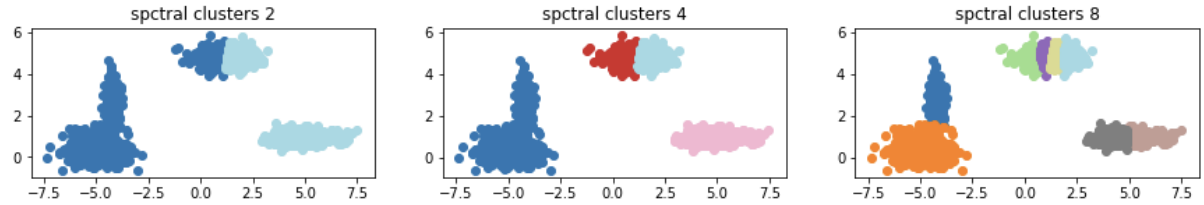
(b)

Spectral Clustering

data 1



data 2



(c)

data 1

Cost comparison on data1

	2 clusters	4 clusters	8 clusters
K-means	5407.931721	4086.668806	2464.152234
Spectral Clustering	5275.376879	4347.133544	2911.885331

data 2

Cost comparison on data2

	2 clusters	4 clusters	8 clusters
K-means	2456.284025	691.023212	483.987016
Spectral Clustering	2461.802947	866.163376	499.610463

- The criterion function is based on the euclidien distance between data in the same group. We can see that K-means outperform spectral clusering for most of the cases. This might because spectral clustering is using a greedy approach when there are more than 2 clusters. And K-means optimizes the clustering based on euclidean distance in a global matter.
- The only case where spectral clustering has a better cost is the 2 clusters case on data1. This might because spectral clustering is also performing a global optimization based on the distance between every data when there is only 2 classes thus getting a smaller cost. On the other hand, K-means has only concerns the distance between the data and their center.

Problem 2.

Answer:

(a).

From question we know

$$\Phi(\mathbf{x}) = K(\cdot, \mathbf{x}) \tag{1}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\Phi(\mathbf{x}_n)\Phi(\mathbf{x}_n)^T \tag{2}$$

$$\hat{\Sigma}\mu = \lambda\mu \tag{3}$$

Inserting 1 into 2, we got

$$\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}K(\cdot, \mathbf{x}_n)K(\cdot, \mathbf{x}_n)^T \tag{4}$$

Inserting 4 into 3, we got

$$\frac{1}{N}\sum_{n=1}^{N}K(\cdot, \mathbf{x}_n)K(\cdot, \mathbf{x}_n)^T\mu = \lambda\mu \tag{5}$$

Assuming that the inner product between $K(\cdot, \mathbf{x}_n)$ and μ are as follow

$$\begin{gathered}\langle K(\cdot, \mathbf{x}_n), \mu\rangle \\ = K(\cdot, \mathbf{x}_n)^T\mu = a_n' \\ \forall n = 1, 2, ..., N\end{gathered} \tag{6}$$

Inserting 6 to 5, we got

$$\begin{gathered}\frac{1}{N}\sum_{n=1}^{N}K(\cdot, \mathbf{x}_n)a_n' = \lambda\mu \\ \frac{1}{\lambda N}\sum_{n=1}^{N}a_n'K(\cdot, \mathbf{x}_n) = \mu\end{gathered} \tag{7}$$

Equation 7 shows that μ is the span of $\{K(\cdot, \mathbf{x}_1), K(\cdot, \mathbf{x}_2), ...K(\cdot, \mathbf{x}_N)\}$

Combining the constant term in 7, we got

$$\mu = \sum_{n=1}^{N}a_nK(\cdot, \mathbf{x}_n) \tag{8}$$

(b)

Combining 5 and 8, we got

$$\frac{1}{N}\sum_{n=1}^{N}K(\cdot, \mathbf{x}_n)K(\cdot, \mathbf{x}_n)^T\sum_{n=1}^{N}a_nK(\cdot, \mathbf{x}_n) = \lambda\sum_{n=1}^{N}a_nK(\cdot, \mathbf{x}_n) \tag{9}$$

Reorganizing 5, we got

$$\frac{1}{N}\sum_{m=1}^{N}(K(\cdot,\mathbf{x}_m)K(\cdot,\mathbf{x}_m)^T\sum_{n=1}^{N}a_nK(\cdot,\mathbf{x}_n))=\lambda\sum_{n=1}^{N}a_nK(\cdot,\mathbf{x}_n)$$
$$\frac{1}{N}\sum_{m=1}^{N}(K(\cdot,\mathbf{x}_m)\sum_{n=1}^{N}a_nK(\cdot,\mathbf{x}_m)^TK(\cdot,\mathbf{x}_n))=\lambda\sum_{n=1}^{N}a_nK(\cdot,\mathbf{x}_n)$$
$$\sum_{m=1}^{N}(K(\cdot,\mathbf{x}_m)\sum_{n=1}^{N}a_nK(\mathbf{x}_m,\mathbf{x}_n))=N\lambda\sum_{n=1}^{N}a_nK(\cdot,\mathbf{x}_n) \tag{10}$$

We know that

$$a=\begin{bmatrix}a_1\\a_2\\\vdots\\a_N\end{bmatrix} \tag{11}$$

Using 11, we can reorganize 10 into

$$\sum_{m=1}^{N}(K(\cdot,\mathbf{x}_m)\begin{bmatrix}K(\mathbf{x}_m,\mathbf{x}_1) & K(\mathbf{x}_m,\mathbf{x}_2) & \dots & K(\mathbf{x}_m,\mathbf{x}_N)\end{bmatrix})a=(\sum_{n=1}^{N}K(\cdot,\mathbf{x}_n))N\lambda a$$

$$\begin{bmatrix}K(\cdot,\mathbf{x}_1) & \dots & K(\cdot,\mathbf{x}_N)\end{bmatrix}\begin{bmatrix}K(\mathbf{x}_1,\mathbf{x}_1) & \dots & K(\mathbf{x}_1,\mathbf{x}_N)\\ \vdots & & \\ K(\mathbf{x}_2,\mathbf{x}_1) & \dots & K(\mathbf{x}_2,\mathbf{x}_N)\\ K(\mathbf{x}_N,\mathbf{x}_1) & \dots & K(\mathbf{x}_N,\mathbf{x}_N)\end{bmatrix}a=\begin{bmatrix}K(\cdot,\mathbf{x}_1) & \dots & K(\cdot,\mathbf{x}_N)\end{bmatrix}N\lambda a$$

$$\begin{bmatrix}K(\mathbf{x}_1,\mathbf{x}_1) & \dots & K(\mathbf{x}_1,\mathbf{x}_N)\\ \vdots & & \\ K(\mathbf{x}_2,\mathbf{x}_1) & \dots & K(\mathbf{x}_2,\mathbf{x}_N)\\ K(\mathbf{x}_N,\mathbf{x}_1) & \dots & K(\mathbf{x}_N,\mathbf{x}_N)\end{bmatrix}a=N\lambda a \tag{12}$$

Define $\mathcal{K}$, we got

$$\mathcal{K}a=N\lambda a \tag{13}$$

(c)

Using equation 8, we have

$$
\begin{aligned}
\langle \mu, \mu \rangle &= \mu^T \mu \\
&= \sum_{n=1}^{N} a_n K(\cdot, \mathbf{x}_n)^T \sum_{n=1}^{N} a_n K(\cdot, \mathbf{x}_n) \\
&= \sum_{m=1}^{N} (a_m \sum_{n=1}^{N} a_n K(\cdot, \mathbf{x}_m)^T K(\cdot, \mathbf{x}_n)) \\
&= \sum_{m=1}^{N} (a_m \sum_{n=1}^{N} a_n K(\mathbf{x}_m, \mathbf{x}_n)) \\
&= \sum_{n=1}^{N} (a_n \begin{bmatrix} K(\mathbf{x}_m, \mathbf{x}_1) & K(\mathbf{x}_m, \mathbf{x}_2) & \dots & K(\mathbf{x}_m, \mathbf{x}_N) \end{bmatrix}) a \\
&= a^T \mathcal{K} a
\end{aligned} \tag{14}
$$

Choose a equals to

$$
a = \frac{z}{\sqrt{v}} \tag{15}
$$

Where z is the eigenvector of $\mathcal{K}$ corresponding to the maximum eigenvalue v

$$
\mathcal{K} z = v z \tag{16}
$$

Inserting 15 into 14, we got

$$
\langle \mu, \mu \rangle = \frac{z^T}{\sqrt{v}} \mathcal{K} \frac{z}{\sqrt{v}} \tag{17}
$$

Inserting 16 into 17, we got

$$
\begin{aligned}
\langle \mu, \mu \rangle &= \frac{z^T}{\sqrt{v}} \frac{v z}{\sqrt{v}} \\
&= z^T z
\end{aligned} \tag{18}
$$

Because z is a normalized unit vector

$$
\langle \mu, \mu \rangle = 1 \tag{19}
$$

(d)

Transform x with the kernel

$$
\mathbf{x} \to \Phi(\mathbf{x}) = K(\cdot, \mathbf{x}) \tag{20}
$$

Projecting x onto kernel principal component μ is to find the inner product of them.

Using 8 and 20, we got

$$
\begin{aligned}
&\langle u, \Phi(\mathbf{x}) \rangle = \mu^T K(\cdot, \mathbf{x}) \\
&= \sum_{n=1}^{N} a_n K(\cdot, \mathbf{x}_n)^T K(\cdot, \mathbf{x}) \\
&= \sum_{n=1}^{N} a_n K(\mathbf{x}_n, \mathbf{x}) \\
&= \sum_{n=1}^{N} a_n K(\mathbf{x}, \mathbf{x}_n)
\end{aligned} \tag{21}
$$

(e)

Problem (b) shows that a can be solved by

$$
\mathcal{K} a = N \lambda a \tag{22}
$$

The ith highest eigenvalue in equation 22 implies ith highest λ in the original problem. This shows that a for the ith kernel principal component is the eigenvector for the ith highest eigenvalue in equation 22.

Problem (c) shows that if we choose a to be eigenvector scaled by the square root of eigenvalue, the kernel principal component can be normalized to have a inner product of one with itself.

Problem (d) shows that the projection of x onto any kernel principal component μ is

$$
\sum_{n=1}^{N} a_n K(\mathbf{x}, \mathbf{x}_n) \tag{23}
$$

Combining the result of problem (b), (c), (d), we find that the projection of x onto the first m kernel principal components is given by

$$
\begin{bmatrix}
\sum_{n=1}^{N} a_n^1 K(\mathbf{x}, \mathbf{x}_n) \\
\sum_{n=1}^{N} a_n^2 K(\mathbf{x}, \mathbf{x}_n) \\
\vdots \\
\sum_{n=1}^{N} a_n^m K(\mathbf{x}, \mathbf{x}_n)
\end{bmatrix} \tag{24}
$$

Where a_n^i is the nth element of ith highest scaled eigenvector a^i

$$
a^i = \frac{z^i}{\sqrt{v^i}} \tag{25}
$$

(f)

Data after centered transformed

$$
\tilde{\Phi}(\mathbf{x}_i) = \Phi(\mathbf{x}_i) - \frac{1}{N} \sum_{n=1}^{N} \Phi(\mathbf{x}_n) \tag{26}
$$

The definition of equation 2 becomes

$$
\tilde{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} \tilde{\Phi}(\mathbf{x}_n) \tilde{\Phi}(\mathbf{x}_n)^T \tag{27}
$$

Inserting 26 into 27, we got

$$\tilde{\Sigma} = \frac{1}{N}\sum_{m=1}^{N}(\Phi(\mathbf{x}_m) - \frac{1}{N}\sum_{n=1}^{N}\Phi(\mathbf{x}_n))(\Phi^T(\mathbf{x}_m) - \frac{1}{N}\sum_{n=1}^{N}\Phi^T(\mathbf{x}_n))$$
$$\hat{\Sigma} = \frac{1}{N}\sum_{m=1}^{N}(\Phi(\mathbf{x}_m)\Phi^T(\mathbf{x}_m) - \frac{1}{N}\sum_{n=1}^{N}\Phi(\mathbf{x}_m)\Phi^T(\mathbf{x}_n)$$
$$-\frac{1}{N}\sum_{n=1}^{N}\Phi(\mathbf{x}_n)\Phi^T(\mathbf{x}_m) + \frac{1}{N^2}\sum_{n=1}^{N}\Phi(\mathbf{x}_n)\sum_{n=1}^{N}\Phi^T(\mathbf{x}_n)) \tag{27}$$

Inserting 1 into 27, we got

$$\tilde{\Sigma} = \frac{1}{N}\sum_{m=1}^{N}(K(\cdot,\mathbf{x}_m)K(\cdot,\mathbf{x}_m)^T - \frac{1}{N}\sum_{n=1}^{N}K(\cdot,\mathbf{x}_m)K(\cdot,\mathbf{x}_n)^T$$
$$-\frac{1}{N}\sum_{n=1}^{N}K(\cdot,\mathbf{x}_n)K(\cdot,\mathbf{x}_m)^T + \frac{1}{N^2}\sum_{n=1}^{N}K(\cdot,\mathbf{x}_n)\sum_{n=1}^{N}K(\cdot,\mathbf{x}_n)^T) \tag{28}$$

Reorganizaing 28, we can see that $\tilde{\Sigma}$ is compose of four parts

$$\tilde{\Sigma} = \frac{1}{N}\sum_{m=1}^{N}K(\cdot,\mathbf{x}_m)K(\cdot,\mathbf{x}_m)^T$$
$$-\frac{1}{N}\sum_{m=1}^{N}(\frac{1}{N}\sum_{n=1}^{N}K(\cdot,\mathbf{x}_m)K(\cdot,\mathbf{x}_n)^T)$$
$$-\frac{1}{N}\sum_{m=1}^{N}(\frac{1}{N}\sum_{n=1}^{N}K(\cdot,\mathbf{x}_n)K(\cdot,\mathbf{x}_m)^T)$$
$$+\frac{1}{N^2}\sum_{n=1}^{N}K(\cdot,\mathbf{x}_n)\sum_{n=1}^{N}K(\cdot,\mathbf{x}_n)^T \tag{29}$$

These 4 components in 29 can be seen as 4 parallel $\Sigma\mu = \lambda\mu$ problem.

And we can follow the steps in problem (b) to solve each of them.

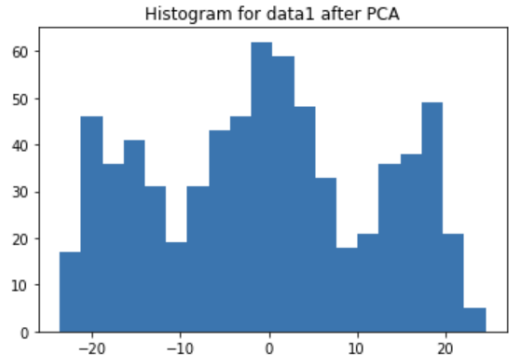
This shows that

$$\tilde{\mathcal{K}} = \mathcal{K} - \frac{1}{N}O\mathcal{K} - \frac{1}{N}\mathcal{K}O + \frac{1}{N^2}O\mathcal{K}O \tag{30}$$

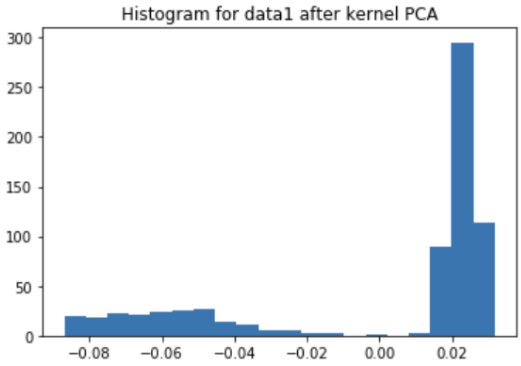
where O is the N x N matrix of all ones.

(g)

Historgram of the projected data1 using PCA



Histogram of the projected data1 using kernel PCA with a Gaussian kernel



The histogram using kernel PCA is has clearer boundary for clustering comparing to the histogram using linear PCA.

Problem 3

Answer:

(a)

If we add a regularization term to the perceptron cost function, the cost function will become

$$J(\theta_0, \vartheta) = -\sum_{n=1}^{N} y_n(\theta_0 + \langle \vartheta, K(\cdot, \mathbf{x}_N)\rangle) + \lambda||\vartheta||^2 \tag{1}$$

We know that ϑ belongs to the RKHS.

We can separate ϑ into two parts.
One parts is in in the span of $\{K(\cdot, \mathbf{x}_1), K(\cdot, \mathbf{x}_2), \ldots \mathbf{x}_N)\}$. And the other part is perpendicular to this span.

$$\vartheta(\cdot) = \sum_{n=1}^{N} \theta_n K(\cdot, \mathbf{x}_n) + \vartheta_\perp \tag{2}$$

From the reproducing property, we know that the first part in the cost in equation 1 can be calculate as follow

$$
\begin{aligned}
&-\sum_{n=1}^{N} y_n(\theta_0 + \langle \vartheta, K(\cdot, \mathbf{x}_N)\rangle) \\
&= -\sum_{n=1}^{N} y_n(\theta_0 + \langle \sum_{m=1}^{N} \theta_m K(\cdot, \mathbf{x}_m) + \vartheta_\perp, K(\cdot, \mathbf{x}_N)\rangle \\
&= -\sum_{n=1}^{N} y_n(\theta_0 + \langle \sum_{m=1}^{N} \theta_m K(\cdot, \mathbf{x}_m), K(\cdot, \mathbf{x}_N)\rangle + \langle \vartheta_\perp, K(\cdot, \mathbf{x}_N)\rangle \\
&= -\sum_{n=1}^{N} y_n(\theta_0 + \langle \sum_{m=1}^{N} \theta_m K(\mathbf{x}_m, \mathbf{x}_N)\rangle \qquad (3)
\end{aligned}
$$

Equation 3 doesn't depends on the perpendicular term.

The second part which is the regularization term in equation 1 can be calculate as follow

$$
\begin{aligned}
\lambda||\vartheta||^2 &= (||\sum_{n=1}^{N} \theta_n K(\cdot, \mathbf{x}_n)||^2 + ||\vartheta_\perp||^2) \\
&\geq ||\sum_{n=1}^{N} \theta_n K(\cdot, \mathbf{x}_n)||^2 \qquad (4)
\end{aligned}
$$

Combining 3 and 4, we know that there exisit $\vartheta(\cdot) = \sum_{n=1}^{N} \theta_n K(\cdot, \mathbf{x}_n)$ that minimize the equation 1.

Using this information, we can define a function g as follow

$$
\begin{aligned}
g(\mathbf{x}) &= \theta_0 + \langle \vartheta, K(\cdot, \mathbf{x})\rangle \\
&= \theta_0 + \vartheta(\mathbf{x}) \\
&= \theta_0 + \sum_{n=1}^{N} \theta_n K(\mathbf{x}, \mathbf{x}_n) \qquad (5)
\end{aligned}
$$

Inserting 5 into the original cost function, we got

$$
J(\mathbf{x}) = -\sum_{n=1}^{N} y_n g(\mathbf{x}) \qquad (6)
$$

Equation 6 shows that g(x) is the function that minimized $J(x)$ and correctly classifies the N training samples.

(b)

ϵ at iteration i-1 is as follow

$$
\epsilon^{(i-1)} = (\theta_0^{(i-1)} - \alpha\theta_0^*) + ||\vartheta^{(i-1)} - \alpha\vartheta^*||_{\mathcal{H}}^2 \qquad (7)
$$

ϵ at iteration i is as follow

$$
\epsilon^{(i)} = (\theta_0^{(i)} - \alpha\theta_0^*) + ||\vartheta^{(i)} - \alpha\vartheta^*||_{\mathcal{H}}^2 \qquad (8)
$$

We know that the update rule when misclassified is as follow

$$
\begin{aligned}
\vartheta^{(i)} &\leftarrow \vartheta^{(i-1)} + \mu y_{(i)} K(\cdot, \mathbf{x}_{(i)}) \\
\theta_0^{(i)} &\leftarrow \theta_0^{(i-1)} + \mu y_{(i)} \qquad (9)
\end{aligned}
$$

Using this update rule in 9, we can rewrite 8 into

$$\begin{aligned}\epsilon^{(i)} = (\theta_0^{(i-1)} + \mu y_{(i)} - \alpha\theta_0^*) \\ +||\vartheta^{(i-1)} + \mu y_{(i)} K(\cdot, \mathbf{x}_{(i)}) - \alpha\vartheta^*||_{\mathcal{H}}^2\end{aligned} \tag{10}$$

Reorganizing 10, we got

$$\begin{aligned}\epsilon^{(i)} &= (\theta_0^{(i-1)} - \alpha\theta_0^*) + ||\vartheta^{(i-1)} - \alpha\vartheta^*||_{\mathcal{H}}^2 \\ &\quad +\mu y_{(i)} + ||\mu y_{(i)} K(\cdot, \mathbf{x}_{(i)})||_{\mathcal{H}}^2 \\ &= \epsilon^{(i-1)} + \mu y_{(i)} + ||\mu y_{(i)} K(\cdot, \mathbf{x}_{(i)})||_{\mathcal{H}}^2 \\ &= \epsilon^{(i-1)} + \mu y_{(i)} + \mu^2 (K(\mathbf{x}_{(i)}, \mathbf{x}_{(i)}))\end{aligned} \tag{11}$$

The prblem want us to show that

$$\epsilon^{(i)} \leq \epsilon^{(i-1)} + \mu^2(1 + K(\mathbf{x}_{(i)}, \mathbf{x}_{(i)})) - 2\alpha\mu y_{(i)} g^*(x_{(i)}) \tag{12}$$

Combining 11 with the problem 12, we got

$$\begin{aligned}\mu y_{(i)} &\leq \mu^2 - 2\alpha\mu y_{(i)} g^*(x_{(i)}) \\ \mu^2 - 2\alpha\mu y_{(i)} g^*(\mathbf{x}_{(i)}) - \mu y_{(i)} &\geq 0\end{aligned} \tag{13}$$

Because $\mu > 0$, equation 13 is true only when

$$\mu \geq 2\alpha y_{(i)} g^*(\mathbf{x}_{(i)}) + y_{(i)} \tag{14}$$

If 14 is true than 12 is true.

(c)

Because $g^*(\mathbf{x}_n)$ correctly classifies every training sample.
Its sign will always be the same as the label y_n.

Therefore

$$\gamma = \min_n y_n g^*(\mathbf{x}_n) > 0 \tag{15}$$

(d)

We know

$$\beta^2 = \max_n K(\mathbf{x}_n, \mathbf{x}_n) \tag{16}$$

$$\alpha = \frac{\mu(1 + \beta^2)}{\gamma} \tag{17}$$

Inserting 17 into 12, we got

$$\begin{aligned}\epsilon^{(i)} &\leq \epsilon^{(i-1)} + \mu^2(1 + K(\mathbf{x}_{(i)}, \mathbf{x}_{(i)})) - 2\frac{\mu(1 + \beta^2)}{\gamma}\mu y_{(i)} g^*(x_{(i)}) \\ \epsilon^{(i)} &\leq \epsilon^{(i-1)} + \mu^2(1 + K(\mathbf{x}_{(i)}, \mathbf{x}_{(i)})) - 2\mu^2(1 + \beta^2)\frac{y_{(i)} g^*(x_{(i)})}{\gamma}\end{aligned} \tag{18}$$

We know that

$$K(\mathbf{x}_{(i)}, \mathbf{x}_{(i)})) \leq \max_n K(\mathbf{x}_n, \mathbf{x}_n) = \beta^2 \tag{19}$$

Using 19 on 18, we got

$$\epsilon^{(i)} \leq \epsilon^{(i-1)} + \mu^2(1+\beta^2)) - 2\mu^2(1+\beta^2)\frac{y_{(i)}g^*(x_{(i)})}{\gamma} \tag{20}$$

We know that

$$y_{(i)}g^*(x_{(i)}) \geq \min_n y_n g^*(\mathbf{x}_n) = \gamma \tag{21}$$

Using 21 on 20, we got

$$\begin{gathered}\epsilon^{(i)} \leq \epsilon^{(i-1)} + \mu^2(1+\beta^2)) - 2\mu^2(1+\beta^2)\frac{\gamma}{\gamma} \\ \epsilon^{(i)} \leq \epsilon^{(i-1)} - \mu^2(1+\beta^2)\end{gathered} \tag{22}$$

(e)

The ϵ at start is as follow

$$\begin{gathered}\epsilon^{(0)} = (\theta_0^{(0)} - \alpha\theta_0^*) + ||\vartheta^{(0)} - \alpha\vartheta^*||_{\mathcal{H}}^2 \\ \epsilon^{(0)} = (0 - \alpha\theta_0^*) + ||0 - \alpha\vartheta^*||_{\mathcal{H}}^2 \\ \epsilon^{(0)} = \alpha^2(\theta_0^*)^2 + \alpha^2||\vartheta^*||_{\mathcal{H}}^2 \\ \epsilon^{(0)} = \alpha^2(\theta_0^*)^2 + \alpha^2\sum_{n=1}^N \theta_n K(\cdot, \mathbf{x}_n)^T \sum_{n=1}^N \theta_n K(\cdot, \mathbf{x}_n) \\ \epsilon^{(0)} = \alpha^2((\theta_0^*)^2 + \sum_{n=1}^N \sum_{m=1}^N \theta_n\theta_m K(\cdot, \mathbf{x}_n)^T K(\cdot, \mathbf{x}_m)) \\ \epsilon^{(0)} = \alpha^2((\theta_0^*)^2 + \sum_{n=1}^N \sum_{m=1}^N \theta_n\theta_m K(\mathbf{x}_n, \mathbf{x}_m))\end{gathered} \tag{23}$$

Inserting α in 17 to 23, we got

$$\epsilon^{(0)} = \frac{\mu^2(1+\beta)^4}{\gamma^2}((\theta_0^*)^2 + \sum_{n=1}^N \sum_{m=1}^N \theta_n\theta_m K(\mathbf{x}_n, \mathbf{x}_m)) \tag{24}$$

From , we know for every iteration ϵ is at least smaller for $\mu^2(1+\beta)^2$ than prevous iteration.

We also know that when $\epsilon = 0$, the algorithm is finished.

Therefore, the iteration is at most

$$\frac{\epsilon^{(0)}}{\mu^2(1+\beta)^2} = \frac{(1+\beta)^2}{\gamma^2}((\theta_0^*)^2 + \sum_{n=1}^N \sum_{m=1}^N \theta_n\theta_m K(\mathbf{x}_n, \mathbf{x}_m))$$

(f) $$\tag{25}$$

From the result in problem (3), we can see that the upper bound of iterations doesn't depend on the step size μ.

However, the step size still affects the length of iterations. From equation 22, we can see that with a larger step size the ϵ will decrease faster.