Problem 1 (True or false)(0.6%)

You don't need to give any explanation for this problem. Just determine the correctness of the following statement and answer "T" for true and "F" for false. The notation here please refer to the course video.

F 1. (0.2 %) The strong duality holds only when the primal problem is convex and satisfies the Slater's conditions.

F 2. (0.2%) The complementary slackness condition, $u_i g_i(x) = 0, \forall i = 1, \ldots m$, in a minimization problem, implies that "whenever $g(\bar{x}) = 0$, then $u_i > 0$".

T 3. (0.2%) The dual function $\theta(\mathbf{u}, \mathbf{v})$ gives a lower bound of the optimal value of the primal problem (as a convex minimization problem in standard form) when $\theta(\mathbf{u}, \mathbf{v}) > -\infty$ and $\mathbf{u} \geq 0$.

Problem 2 (SVM with Gaussian kernel)(0.9%)

Consider the task of training a support vector machine using the Gaussian kernel $K(x, z) = \exp(-\frac{\|x-z\|^2}{\tau^2})$. We will show that as long as there are no two identical points in the training set, we can always find a value for the bandwidth parameter τ such that the SVM achieves zero training error.

Recall from class that the decision function learned by the support vector machine can be written as

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + b$$

Assume that the training data $\{(x_1, y_1), \cdots, (x_N, y_N)\}$ consists of points which are separated by at least distance of ϵ; that is, $\|x_j - x_i\| \geq \epsilon$, for any $i \neq j$. For simplicity, we assume $\alpha_i = 1$ for all $i = 1, \cdots, m$ and $b = 0$. Find values for the Gaussian kernel width τ such that x_i is correctly classified, for all $i = 1, \cdots, N$, e.g., $f(x_i)y_i > 0$ for all $i = 1, \cdots, N$.

Hint: Notice that for $y \in \{-1, +1\}$ the prediction on x_i will be correct if $|f(x_i) - y_i| < 1$, so find a value of τ that satisfies this inequality for all i.

The decision function learned by the support vector machine:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \xrightarrow{a_i = 1 \text{ 且 } b=0} f(x) = \sum_{i=1}^{N} y_i K(x_i, x)$$

$$f(x_i) = \sum_{j=1}^{N} y_j k(x_j, x_i) \Rightarrow f(x_i) = y_i K(x_i, x_i) + \sum_{j \neq i} y_j k(x_j, x_i)$$

$$K(x, z) = \exp\left(-\frac{\|x-z\|^2}{\tau^2}\right) \quad \begin{cases} x = z, \ k(x_i, x_i) = \exp(0) = 1 \\ x \neq z \text{ 且 } \|x - z\| \geq \epsilon, \ k(x_j, x_i) \leq \exp(-\frac{\epsilon^2}{\tau^2}) \end{cases}$$

$$f(x_i) = y_i \times 1 + \sum_{j \neq i} y_i k(x_j, x_i)$$

$$E_i = |f(x_i) - y_i| = \left|\sum_{j \neq i} y_j k(x_j, x_i)\right|$$

$$E_i \leq (N-1) \exp\left(-\frac{\epsilon^2}{\tau^2}\right)$$

$$|f(x_i) - y_i| < 1 \xrightarrow{\text{代入误差}} (N-1)\exp\left(-\frac{\epsilon^2}{\tau^2}\right) < 1$$

$$\ln\left((N-1)\exp\left(-\frac{\epsilon^2}{\tau^2}\right)\right) < \ln(1) \rightarrow \ln(N-1) - \frac{\epsilon^2}{\tau^2} < 0$$

$$-\frac{\epsilon^2}{\tau^2} < -\ln(N-1) \Rightarrow \tau^2 < \frac{\epsilon^2}{\ln(N-1)}$$

$$\Rightarrow \because \tau > 0 \therefore \tau < \frac{\epsilon}{\sqrt{\ln(N-1)}}$$

Problem 3 (Support Vector Regression)(1.5%)

Suppose we are given a training set $\{(x_1, y_1), \cdots, (x_m, y_m)\}$, where $x_i \in \mathbb{R}^{(n+1)}$ and $y_i \in \mathbb{R}$. We would like to find a hypothesis of the form $f(x) = w^T x + b$. It is possible that no such function $f(x)$ exists to satisfy these constraints for all points. To deal with otherwise infeasible constraints, we introduce slack variables ξ_i for each point. The (convex) optimization problem is

$$
\begin{aligned}
&\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i && && (1)\\
&\text{s.t. } y_i - w^T x_i - b \le \epsilon + \xi_i && i = 1, \dots, m && (2)\\
&\quad\;\; w^T x_i + b - y_i \le \epsilon + \xi_i && i = 1, \dots, m && (3)\\
&\quad\;\; \xi_i \ge 0 && i = 1, \dots, m && (4)
\end{aligned}
$$

where $\epsilon > 0$ is a given, fixed value and $C > 0$. Denote that $\xi = (\xi_1, \cdots, \xi_m)$.

(a) Write down the Lagrangian for the optimization problem above. Consider the sets of Lagrange multiplier α_i, α_i^*, β_i corresponding to the (2), (3), and (4), so that the Lagrangian would be written as $\mathcal{L}(w, b, \xi, \alpha, \alpha^*, \beta)$, where $\alpha = (\alpha_1, \cdots, \alpha_m)$, $\alpha^* = (\alpha_1^*, \cdots, \alpha_m^*)$, and $\beta = (\beta_1, \cdots, \beta_m)$.

(b) Derive the dual optimization problem. You will have to take derivatives of the Lagrangian with respect to w, b, and ξ

(a) $\mathcal{L}(w,b,\xi,\alpha,\alpha^*,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - (y_i - w^T x_i - b)) + \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i - (w^T x_i + b - y_i)) - \sum_{i=1}^{m}\beta_i\xi_i$

$\alpha = (\alpha_1, \dots, \alpha_m)$, $\alpha^*(\alpha_1^*, \dots \alpha_m^*)$, $\beta = (\beta_1, \dots, \beta_m)$

(b) $\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{m}\alpha_i x_i + \sum_{i=1}^{m}\alpha_i^* x_i = 0 \Rightarrow w = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_i$

$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{m}\alpha_i + \sum_{i=1}^{m}\alpha_i^* = 0 \Rightarrow \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$

$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \alpha_i^* - \beta_i = 0$

$\Rightarrow \alpha_i + \alpha_i^* + \beta_i = C$

$\max_{\alpha,\alpha^*}\left(-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^T x_j + \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)y_i - \varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*)\right)$

$\sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$, $0 \le \alpha_i \le C$ & $0 \le \alpha_i^* \le C$.

(c) Suppose that $(\bar{w}, \bar{b}, \bar{\xi})$ and $(\bar{\alpha}, \bar{\alpha^*}, \bar{\beta})$ are the optimal solutions to a primal and dual optimization problem, respectively.

Denote $\bar{w} = \sum_{i=1}^m (\bar{\alpha}_i - \bar{\alpha}_i^*) x_i$

(1) Prove that

$$\bar{b} = \arg\min_{b \in \mathbb{R}} C \sum_{i=1}^m \max(|y_i - (\bar{w}^T x_i + b)| - \epsilon, 0) \tag{5}$$

(2) Define $e = y_i - (\overline{w}^T x_i + \bar{b})$ Prove that

$$\begin{cases} \bar{\alpha}_i = \bar{\alpha}_i^* = 0, & \bar{\xi}_i = 0, & \text{if } |e| < \epsilon \\ 0 \le \bar{\alpha}_i \le C, & \bar{\xi}_i = 0, & \text{if } e = \epsilon \\ 0 \le \bar{\alpha}_i^* \le C, & \bar{\xi}_i = 0, & \text{if } e = -\epsilon \\ \bar{\alpha}_i = C, & \bar{\xi}_i = e - \epsilon & \text{if } e > \epsilon \\ \bar{\alpha}_i^* = C, & \bar{\xi}_i = -(e + \epsilon) & \text{if } e < -\epsilon \end{cases} \tag{6}$$

(C) ① $\xi_i \geq |y_i - (w^T x_i + b)| - \varepsilon$

$\xi_i = \max(|y_i - (w^T x_i + b)| - \varepsilon, 0)$

$w = \bar{w} \Rightarrow \xi_i = \max(|y_i - (\bar{w}^T x_i + b)| - \varepsilon, 0)$

最小化 $C \sum_{i=1}^m \xi_i$

$\bar{b} = \arg\min_{b \in R} C \sum_{i=1}^m \max(|y_i - (\bar{w}^T x_i + b)| - \varepsilon, 0)$

② Case $|e| < \varepsilon$:

The error is within the margin, so no penalty is incurred. $\therefore \xi_i = 0$

$\because$ no penalty $\Rightarrow \bar{\alpha}_i = \bar{\alpha}_i^* = 0$

Case $e = \varepsilon$:

The prediction is exactly on the upper margin $\Rightarrow \xi_i = 0$

$\because$ any value in this range would satisfy the KKT conditions $\therefore 0 \le \bar{\alpha}_i \le C$

Case $e = -\varepsilon$:

The prediction is on the lower margin. $\Rightarrow \xi_i = 0$

$\because$ any value in this range would satisfy the KKT conditions $\therefore 0 \le \bar{\alpha}_i^* \le C$

Case $e > \varepsilon$:

The error exceeds the upper margin $\therefore$ there is a positive slack penalty $\xi_i = e - \varepsilon$

Satisfy the KKT conditions, $\bar{\alpha}_i = C$, indicating that the point is penalized to its maximum extent.

Case $e < -\varepsilon$:

the error exceeds the lower margin, $\therefore$ there is a positive slack penalty $\xi_i = -(e + \varepsilon)$

Satisfy the KKT conditions, $\bar{\alpha}_i^* = C$, indicating that the point is penalized to its maximum extent.

(d) Show that the algorithm can be kernelized and write down the kernel form of the decision function. For this, you have to show that

(1) The dual optimization objective can be written in terms of inner products or training examples

(2) At test time, given a new x the hypothesis $f(x)$ can also be computed in terms of inner produce.

(d) ① $\max_{\alpha,\alpha^*} \left(-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)x_i^T x_j+\sum_{i=1}^{m}(\alpha_i-\alpha_i^*)y_i-\epsilon\sum_{i=1}^{m}(\alpha_i+\alpha_i^*)\right)$

subject to: $\sum_{i=1}^{m}(\alpha_i-\alpha_i^*)=0$. $0\le\alpha_i,\ \alpha_i^*\le C\quad\forall i$

The dual objective function $W(\alpha,\alpha^*)$ depends on the data x_i only through the inner products $x_i^T x_j$.

The variables α_i and α_i^* are the dual variables associated with the constraints, and y_i are the target values.

$x_i^T x_j$ represents the inner product between training examples x_i and x_j

By substituting $x_i^T x_j$ with a kernel function $K(x_i,x_j)$, we can compute the inner products in a transformed feature space.

Replace $x_i^T x_j$ with $K(x_i,x_j)$:

$\max_{\alpha,\alpha^*} W(\alpha,\alpha^*)=-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j)+\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)y_i-\epsilon\sum_{i=1}^{m}(\alpha_i+\alpha_i^*)$

subject to: $\sum_{i=1}^{m}(\alpha_i-\alpha_i^*)=0\quad 0\le\alpha_i,\alpha_i^*\le C\quad\forall i$

② decision function in the primal form: $f(x)=w^T x+b$

w in terms of the dual variables: $w=\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)x_i$

$\Rightarrow f(x)=\left(\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)x_i\right)^T x+b=\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)x_i^T x+b$

Replace $x_i^T x$ with $K(x_i,x)$: $f(x)=\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)K(x_i,x)+b$

The $f(x)$ depends only on the kernel evaluations $K(x_i,x)$ between the test point x and the training examples x_i.

$(\alpha_i^*-\alpha_i)$ are determined during training by solving the dual optimization problem.

The bias term b can be computed using the KKT conditions, often involving support vectors for which

$0<\alpha_i<C$ or $0<\alpha_i^*<C$

Problem 4 (Hinge loss with L^1 regularization)(1.5%)

Given data points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ as well as their labels $y_1, \ldots, y_N \in \{\pm 1\}$ and penalty coefficients $C_1, \ldots, C_N > 0$, where each $\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,m}]^T$ is a column vector, consider the following optimization problem:

$$
\begin{array}{ll}
\text{minimize} & \|\mathbf{w}\|_1 + \sum_{i=1}^N C_i \xi_i \\
\text{subject to} & \left.\begin{array}{l} y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{array}\right\} \; i = 1, \ldots, N \\
\text{variables} & \mathbf{w} \in \mathbb{R}^m, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N
\end{array}
\tag{7}
$$

Note that in this formulation, we replace the L^2-regularization term $\frac{1}{2}\|\mathbf{w}\|^2$ by the L^1-regularization term $\|\mathbf{w}\|_1 = \sum_{j=1}^m |w_j|$.

(a) (0.5%) Show that $(\bar{\mathbf{w}}, \bar{b}, \bar{\boldsymbol{\xi}})$ is an optimal solution of (7) if and only if $\bar{\mathbf{w}} = \bar{\mathbf{u}} - \bar{\mathbf{v}}$, where $(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \bar{b}, \bar{\boldsymbol{\xi}})$ is an optimal solution of the following problem:

$$
\begin{array}{ll}
\text{minimize} & f(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}) = \sum_{j=1}^m (u_j + v_j) + \sum_{i=1}^N C_i \xi_i \\
\text{subject to} & y_i((\mathbf{u} - \mathbf{v})^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, N \\
 & \xi_i \geq 0, \quad i = 1, \ldots, N \\
 & u_j \geq 0, \quad v_j \geq 0, \quad j = 1, \ldots, m \\
\text{variables} & \mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^m, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N
\end{array}
\tag{8}
$$

Following (a), we can now rewrite (8) as the following primal problem:

$$
\begin{array}{ll}
\text{minimize} & f(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}) = \sum_{j=1}^m (u_j + v_j) + \sum_{i=1}^N C_i \xi_i \\
\text{subject to} & g_i^1(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}) = 1 - \xi_i - y_i((\mathbf{u} - \mathbf{v})^T \mathbf{x}_i + b) \leq 0, \quad i = 1, \ldots, N \\
 & g_i^2(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}) = -\xi_i \leq 0, \quad i = 1, \ldots, N \\
 & g_i^3(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}) = -u_j \leq 0, \quad j = 1, \ldots, m \\
 & g_i^4(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}) = -v_j \leq 0, \quad j = 1, \ldots, m \\
\text{variables} & \mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^m, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N
\end{array}
\tag{9}
$$

as well as its Lagrangian dual problem:

$$
\begin{array}{ll}
\text{maximize} & \theta(\alpha, \beta, \boldsymbol{\mu}, \boldsymbol{\nu}) = \inf\left\{L(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}, \alpha, \beta, \boldsymbol{\mu}, \boldsymbol{\nu}) : \mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^m, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N\right\} \\
\text{subject to} & \alpha_i \geq 0, \beta_i \geq 0 \quad, i = 1, \ldots, N \\
 & \mu_j \geq 0, \nu_j \geq 0 \quad, j = 1, \ldots, m \\
\text{variables} & \alpha \in \mathbb{R}^N, \beta \in \mathbb{R}^N, \boldsymbol{\mu} \in \mathbb{R}^m, \boldsymbol{\nu} \in \mathbb{R}^m
\end{array}
\tag{10}
$$

where L denotes the Lagranian function.

(a) from (7) to (8)

For each component $j \Rightarrow \bar{u}_j = \max\{\bar{w}_j, 0\}$, $\bar{v}_j = \max\{-\bar{w}_j, 0\}$ $\Rightarrow \bar{u}_j \geq 0, \bar{v}_j \geq 0, \bar{w}_j = \bar{u}_j - \bar{v}_j$

$\|\bar{w}\|_1 = \sum_{j=1}^m |\bar{w}_j| = \sum_{j=1}^m (\bar{u}_j + \bar{v}_j) \xrightarrow{\text{in (7)}} \|\bar{w}\|_1 + \sum_{i=1}^N C_i \bar{\xi}_i = \sum_{j=1}^m (\bar{u}_j + \bar{v}_j) + \sum_{i=1}^N C_i \bar{\xi}_i$ (7)和(8)的目標函數一致

The constraints in (7): $y_i(\bar{w}^T x_i + \bar{b}) \geq 1 - \bar{\xi}_i$, $\bar{\xi}_i \geq 0 \xrightarrow{\bar{w} = \bar{u} - \bar{v}} y_i((\bar{u} - \bar{v})^T x_i + \bar{b}) \geq 1 - \bar{\xi}_i$, $\bar{\xi}_i \geq 0$

$\because$ $(\bar{w}, \bar{b}, \bar{\xi})$ 是(7)的最佳解，且 $(\bar{u}, \bar{v}, \bar{b}, \bar{\xi})$ 在(8)中得相同的目標函數值並滿足約束條件 $\therefore$ $(\bar{u}, \bar{v}, \bar{b}, \bar{\xi})$ 是8的最佳解

from (8) to (7)

define $\bar{w} = \bar{u} - \bar{v} \xrightarrow{\text{L1-norm of } \bar{w}} \|\bar{w}\|_1 = \sum_{j=1}^m |\bar{w}_j| = \sum_{j=1}^m (\bar{u}_j + \bar{v}_j) \xrightarrow{\text{objective function in (8)}} \sum_{j=1}^m (\bar{u}_j + \bar{v}_j) + \sum_{i=1}^N C_i \bar{\xi}_i$ (matches (7) when expressed in terms of $\bar{w}$)

The constraints in (8): $y_i((\bar{u} - \bar{v})^T x_i + \bar{b}) \geq 1 - \bar{\xi}_i$, $\bar{\xi}_i \geq 0 \Rightarrow y_i(\bar{w}^T x_i + \bar{b}) \geq 1 - \bar{\xi}_i$, $\bar{\xi}_i \geq 0$ (正是(7)的約束條件)

$\because$ $(\bar{u}, \bar{v}, \bar{b}, \bar{\xi})$ 是(8)的最佳解，且 $(\bar{w}, \bar{b}, \bar{\xi})$ 在(7)中得相同的目標函數並滿足約束條件，$\therefore$ $(\bar{w}, \bar{b}, \bar{\xi})$ 是(7)的最佳解

(b) (0.2%) Associate dual variables $\alpha_i, \beta_i, \mu_j, \nu_j$ to constraints $g_i^1, g_i^2, g_j^3, g_j^4$, respectively. Show that L can be written in the following explicit form:

$$L(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}, \alpha, \beta, \boldsymbol{\mu}, \boldsymbol{\nu}) = \mathbf{1}^T(\mathbf{u}+\mathbf{v}) + \sum_{i=1}^{N} C_i\xi_i + \sum_{i=1}^{N} \alpha_i \left(1 - \xi_i - y_i((\mathbf{u}-\mathbf{v})^T\mathbf{x}_i + b)\right) + \sum_{i=1}^{N} \beta_i(-\xi_i) - \boldsymbol{\mu}^T\mathbf{u} - \boldsymbol{\nu}^T\mathbf{v}, \tag{11}$$

where $\mathbf{1}$ denotes the all-one vector.

(c) (0.1%) Show that (9) satisfies Slater's condition.

(b) $\alpha_i \geq 0$ to the constraints $g_i^1(u,v,b,\xi) = 1 - \xi_i - y_i((u-v)^T x_i + b) \leq 0$

$\beta_i \geq 0$ to the constraints $g_i^2(u,v,b,\xi) = -\xi_i \leq 0$

$\mu_j \geq 0$ to the constraints $g_j^3(u,v,b,\xi) = -u_j \leq 0$

$v_j \geq 0$ to the constraints $g_j^4(u,v,b,\xi) = -v_j \leq 0$

$\mathcal{L}(u,v,b,\xi,\alpha,\beta,\mu,\nu) = f(u,v,b,\xi) + \sum_{i=1}^{N}\alpha_i g_i^1(u,v,b,\xi) + \sum_{i=1}^{N}\beta_i g_i^2(u,v,b,\xi) + \sum_{j=1}^{m}\mu_j g_j^3(u,v,b,\xi) + \sum_{j=1}^{m}\nu_j g_j^4(u,v,b,\xi)$

substituting $f(u,v,b,\xi)$ and the constraints:

$\mathcal{L}(u,v,b,\xi,\alpha,\beta,\mu,\nu) = \sum_{j=1}^{m}(u_j+v_j) + \sum_{i=1}^{N} C_i\xi_i + \sum_{i=1}^{N}\alpha_i(1-\xi_i-y_i((u-v)^T x_i+b)) + \sum_{i=1}^{N}\beta_i(-\xi_i) + \sum_{j=1}^{m}\mu_j(-u_j) + \sum_{j=1}^{m}\nu_j(-v_j)$

$\Rightarrow \mathcal{L}(u,v,b,\xi,\alpha,\beta,\mu,\nu) = \sum_{j=1}^{m}(u_j+v_j) + \sum_{i=1}^{N} C_i\xi_i + \sum_{i=1}^{N}\alpha_i(1-\xi_i-y_i((u-v)^T x_i+b)) - \sum_{i=1}^{N}\beta_i\xi_i - \sum_{j=1}^{m}\mu_j u_j - \sum_{j=1}^{m}\nu_j v_j$

Expressed in vector notation, using 1 as the all-one vector:

$\mathcal{L}(u,v,b,\xi,\alpha,\beta,\mu,\nu) = 1^T(u+v) + \sum_{i=1}^{N} C_i\xi_i + \sum_{i=1}^{N}\alpha_i(1-\xi_i-y_i((u-v)^T x_i+b)) - \sum_{i=1}^{N}\beta_i\xi_i - \mu^T u - \nu^T v$

(C) let $u_j^\circ = 1$ and $v_j^\circ = 1$ for all $j = 1, \dots, m$. $\Rightarrow$ ensure $u_j^\circ > 0$ and $v_j^\circ > 0$

Set $\xi_i^\circ = 2$ for all $i = 1, \dots, N$ $\Rightarrow$ ensure $\xi_i^\circ > 0$

$b^\circ = 0$

(1) $g_i^1(u^\circ, v^\circ, b^\circ, \xi^\circ) = 1 - \xi_i^\circ - y_i((u^\circ - v^\circ)^T x_i + b^\circ)$ $\because u^\circ = v^\circ$, $(u^\circ - v^\circ)^T x_i = 0$ $\therefore g_i^1 = 1 - 2 - 0 = -1 < 0$

(2) $g_i^2(u^\circ, v^\circ, b^\circ, \xi^\circ) = -\xi_i^\circ = -2 < 0$

(3) $g_j^3(u^\circ, v^\circ, b^\circ, \xi^\circ) = -u_j^\circ = -1 < 0$

(4) $g_j^4(u^\circ, v^\circ, b^\circ, \xi^\circ) = -v_j^\circ = -1 < 0$

$\because$ we have found a point $(u^\circ, v^\circ, b^\circ, \xi^\circ)$ where all inequality contraints are strictly less than 0, Slater's condition is satisfied for (9)

(d) (0.2%) Show that:

(i) (0.1%) $\theta(\alpha, \beta, \boldsymbol{\mu}, \boldsymbol{\nu}) = -\infty$ unless the following conditions hold:

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad \boldsymbol{\mu} = \mathbf{1} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i, \quad \boldsymbol{\nu} = \mathbf{1} + \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i, \tag{12}$$

and

$$\alpha_i + \beta_i = C_i, \quad \forall i = 1, \ldots, N, \tag{13}$$

at which case $\theta(\alpha, \beta, \boldsymbol{\mu}, \boldsymbol{\nu}) = \sum_{i=1}^{N} \alpha_i$.

(ii) (0.1%) The stationary condition holds if and only if (12) and (13) are satisfied.

(d) (i) for μ: $1^T - \mu^T - \sum_{i=1}^{N} \alpha_i y_i x_i^T = 0^T \Rightarrow \mu = 1 - \sum_{i=1}^{N} \alpha_i y_i x_i$

for ν: $1^T - \nu^T + \sum_{i=1}^{N} \alpha_i y_i x_i^T = 0^T \Rightarrow \nu = 1 + \sum_{i=1}^{N} \alpha_i y_i x_i$

The term involving b in L: $-\sum_{i=1}^{N} \alpha_i y_i b$

$-\sum_{i=1}^{N} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0$

The term involving ξ are: $\sum_{i=1}^{N} (C_i \xi_i - \alpha_i \xi_i - \beta_i \xi_i) = \sum_{i=1}^{N} (C_i - \alpha_i - \beta_i) \xi_i$

$\because \xi_i \geq 0$, to prevent L from being unbounded below, $\Rightarrow C_i - \alpha_i - \beta_i = 0 \Rightarrow \alpha_i - \beta_i = C_i$, $\forall i = 1, \ldots, N$

Equality Constraints: $\sum_{i=1}^{N} \alpha_i y_i = 0$, $\alpha_i + \beta_i = C_i$, $\forall i = 1, \ldots, N$

Expressions for μ and ν: $\mu = 1 - \sum_{i=1}^{N} \alpha_i y_i x_i$, $\nu = 1 + \sum_{i=1}^{N} \alpha_i y_i x_i$

the infimum of $\mathcal{L}$ is unbounded below, $\theta(\alpha, \beta, \mu, \nu) = -\infty$

$\mathcal{L}(u, v, b, \xi, \alpha, \beta, \mu, \nu) = \sum_{i=1}^{N} \alpha_i$ $\because \mathcal{L}$ is constant with respect to u, v, b, and ξ when the conditions are satisfied, the infimum is simply: $\theta(\alpha, \beta, \mu, \nu) = \sum_{i=1}^{N} \alpha_i$

(ii) $\nabla_u L = 0 \Rightarrow 1 - \mu - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \Rightarrow \mu = 1 - \sum_{i=1}^{N} \alpha_i y_i x_i$

$\nabla_v L = 0 \Rightarrow 1 - \nu + \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \Rightarrow \nu = 1 + \sum_{i=1}^{N} \alpha_i y_i x_i$

$\frac{\partial L}{\partial b} = 0 \Rightarrow -\sum_{i=1}^{N} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0$

$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C_i - \alpha_i - \beta_i = 0 \Rightarrow \alpha_i + \beta_i = C_i$

The stationary conditions hold if and only if the following are satisfied

$\sum_{i=1}^{N} \alpha_i y_i = 0$, $\alpha_i + \beta_i = C_i$, $\forall i = 1, \ldots, N$

(e) (0.2%) Show that the dual problem (10) can be simplified as:

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^N \alpha_i \\ \text{subject to} & \sum_{i=1}^N \alpha_i y_i = 0 \\ & -\mathbf{1} \le \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \le \mathbf{1} \\ \text{variables} & 0 \le \alpha_i \le C_i, \quad i = 1, \dots, N \end{array} \tag{14}$$

(f) (0.2%) Write down the KKT conditions for the primal and dual problems (9)(10).

(g) (0.1%) Is it true that $\bar{\mathbf{w}}$ must be a linear combination of $\mathbf{x}_1, \dots, \mathbf{x}_N$? Justify your answer.

(e) maximize $\theta(\alpha, \beta, \mu, \nu) = \sum_{i=1}^N \alpha_i$

subject to $\sum_{i=1}^N \alpha_i y_i = 0$, $\mu = 1 - \sum_{i=1}^N \alpha_i y_i x_i \ge 0$, $\nu = 1 + \sum_{i=1}^N \alpha_i y_i x_i \ge 0$, $\alpha_i + \beta_i = C_i$, $\alpha_i \ge 0, \beta_i \ge 0, \forall i = 1, \dots, N$, $\mu_j \ge 0, \nu_j \ge 0$ $\forall j = 1, \dots, m$

Eliminate β_i: $\because \alpha_i + \beta_i = C_i$ and $\beta_i \ge 0 \Rightarrow 0 \le \alpha_i \le C_i$

Express μ and ν:

$$\begin{cases} \mu = 1 - \sum_{i=1}^N \alpha_i y_i x_i & \mu \ge 0 \Rightarrow 1 - \sum_{i=1}^N \alpha_i y_i x_i \ge 0 \Rightarrow \sum_{i=1}^N \alpha_i y_i x_i \le 1 \\ \nu = 1 + \sum_{i=1}^N \alpha_i y_i x_i & \nu \ge 0 \Rightarrow 1 + \sum_{i=1}^N \alpha_i y_i x_i \ge 0 \Rightarrow \sum_{i=1}^N \alpha_i y_i x_i \ge -1 \end{cases} \Rightarrow -1 \le \sum_{i=1}^N \alpha_i y_i x_i \le 1$$

simplified dual problem $\Rightarrow$ maximize $\sum_{i=1}^N \alpha_i$

subject to $\sum_{i=1}^N \alpha_i y_i = 0$

$-1 \le \sum_{i=1}^N \alpha_i y_i x_i \le 1$

$0 \le \alpha_i \le C_i$, $\forall i = 1, \dots, N$

(f) For the primal problem (9)

Primal variables: $u \in \mathbb{R}^m$, $v \in \mathbb{R}^m$, $b \in \mathbb{R}$, $\xi \in \mathbb{R}^N$

$$\begin{cases} y_i((u-v)^T x_i + b) \ge 1 - \xi_i, & i = 1, \dots, N \\ \xi_i \ge 0, & i = 1, \dots, N \\ u_j \ge 0, \ v_j \ge 0, & j = 1, \dots, m \end{cases}$$

Associated Dual Variables:

$\alpha_i \ge 0$ for the first constraint (g_i^1)

$\beta_i \ge 0$ for $\xi_i \ge 0$ (g_i^2)

$\mu_j \ge 0$ for $u_j \ge 0$ (g_j^3)

$\nu_j \ge 0$ for $v_j \ge 0$ (g_j^4)

primal Feasibility:

$$\begin{cases} y_i((u-v)^T x_i + b) \ge 1 - \xi_i, \ \xi_i \ge 0 \\ u_j \ge 0, \ v_j \ge 0 \end{cases}$$

Dual Feasibility: $\alpha_i \ge 0$, $\beta_i \ge 0$, $\mu_j \ge 0$, $\nu_j \ge 0$

Complementary Slackness:

$$\begin{cases} \alpha_i (y_i((u-v)^T x_i + b) - 1 + \xi_i) = 0, \ \beta_i \xi_i = 0 \\ \mu_j u_j = 0, \ \nu_j v_j = 0 \end{cases}$$

$\frac{\partial L}{\partial u} = 1 - \mu - \sum_{i=1}^N \alpha_i y_i x_i = 0$

$\frac{\partial L}{\partial v} = 1 - \nu + \sum_{i=1}^N \alpha_i y_i x_i = 0$

$\frac{\partial L}{\partial b} = - \sum_{i=1}^N \alpha_i y_i = 0$

$\frac{\partial L}{\partial \xi_i} = C_i - \alpha_i - \beta_i = 0$

for the Dual problem (10):

Dual variables: α_i

Constraints:
$$\begin{cases} \sum_{i=1}^N \alpha_i y_i = 0 \\ -1 \le \sum_{i=1}^N \alpha_i y_i x_i \le 1 \\ 0 \le \alpha_i \le C_i \end{cases}$$

KKT:

Dual Feasibility: $0 \le \alpha_i \le C_i$

Primal Feasibility: $\sum_{i=1}^N \alpha_i y_i = 0$, $-1 \le \sum_{i=1}^N \alpha_i y_i x_i \le 1$

Complementary Slackness:

$\alpha_i (y_i((u-v)^T x_i + b) - 1 + \xi_i) = 0$

$(C_i - \alpha_i)\xi_i = 0$

(g) No, $\bar{w}$ does not necessarily have to be a linear combination of $x_1, x_2, \dots, x_N$, when using L_1 regularization.

In L_2 regularized support vector machine, the optimal weight vector $\bar{w}$ can be expressed as a linear combination of the training data vectors due to the representer theorem. $\Rightarrow \bar{w} = \sum_{i=1}^N \alpha_i y_i x_i$

L_1-regularization is different:

① The L_1 norm promotes sparsity in the weight vector $\bar{w}$, meaning many components of $\bar{w}$ may be 0

② The dual variables α_i in the L_1 regularized problem do not directly define $\bar{w}$ as they do in the L_2 regularized case.

③ The optimal $\bar{w}$ arises from solving the primal problem, which may not yield a representation of $\bar{w}$ as a combination of the data points.

Problem 5 (Spherical one class SVM)(1.5%)

Suppose we aim to fit a hypersphere which encompasses a majority of data points $\mathbf{x}_1, ..., \mathbf{x}_N \in \mathbb{R}^M$ by considering the following optimization problem: (here $\boldsymbol{\mu}$ and each $\mathbf{x}_i$ are considered as column vectors)

$$\begin{array}{ll} \text{minimize} & R^2 + \frac{1}{\nu}\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & \left.\begin{array}{l} \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \le R^2 + \xi_i \\ \xi_i \ge 0 \end{array}\right\} \forall i \in [1, N] \\ & R \ge 0 \\ \text{variables} & R \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} = (\xi_1, ..., \xi_N) \in \mathbb{R}^N \end{array} \tag{15}$$

where $C_i > 0$ for each $i \in [1, N]$, and $0 < \nu < \sum_{i=1}^N C_i$. Let $\rho = R^2$ and rewrite (15) in the form of primal problem:

$$\begin{array}{ll} \text{minimize} & f(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = \rho + \frac{1}{\nu}\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & \left.\begin{array}{l} g_{1,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 - \rho - \xi_i \le 0 \\ g_{2,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = -\xi_i \le 0 \end{array}\right\} \forall i \in [1, N] \\ & g_3(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = -\rho \le 0 \\ \text{variables} & \rho \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} \in \mathbb{R}^N \end{array} \tag{16}$$

as well its Lagrangian dual problem:

$$\begin{array}{ll} \text{maximize} & \theta(\alpha, \beta, \gamma) = \inf_{\rho \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} \in \mathbb{R}^N} L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma) \\ \text{subject to} & \alpha_i \ge 0, \beta_i \ge 0 \ \forall i \in [1, N] \\ & \gamma \ge 0 \\ \text{variables} & \alpha = (\alpha_1, ..., \alpha_N) \in \mathbb{R}^N, \beta = (\beta_1, ..., \beta_N) \in \mathbb{R}^N, \gamma \in \mathbb{R} \end{array} \tag{17}$$

1. Write down the Lagrangian function $L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma)$ in explicit form of $\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma$.
2. Show that the duality gap between (16) and (17) is zero.
3. Derive $\theta(\alpha, \beta, \gamma)$ in explicit form of dual variables α, β, γ.

1. $\mathcal{L}(\rho, \mu, \xi, \alpha, \beta, \gamma) = f(\rho, \mu, \xi) + \sum_{i=1}^{N} \alpha_i g_{1,i}(\rho, \mu, \xi) + \sum_{i=1}^{N} \beta_i g_{2,i}(\rho, \mu, \xi) + \gamma g_3(\rho, \mu, \xi)$

$= \rho + \frac{1}{\nu}\sum_{i=1}^{N} C_i \xi_i + \sum_{i=1}^{N} \alpha_i (\|x_i - \mu\|^2 - \rho - \xi_i) + \sum_{i=1}^{N} \beta_i(-\xi_i) + \gamma(-\rho)$

$= \rho(1 - \sum_{i=1}^{N}\alpha_i - \gamma) + \sum_{i=1}^{N} \xi_i \left(\frac{C_i}{\nu} - \alpha_i - \beta_i\right) + \sum_{i=1}^{N} \alpha_i \|x_i - \mu\|^2$

the explicit form of Lagrangian function :

$\mathcal{L}(\rho, \mu, \xi, \alpha, \beta, \gamma) = \rho(1 - \sum_{i=1}^{N}\alpha_i - \gamma) + \sum_{i=1}^{N} \xi_i \left(\frac{C_i}{\nu} - \alpha_i - \beta_i\right) + \sum_{i=1}^{N} \alpha_i \|x_i - \mu\|^2$

2. primal problem:

objective function convexity: $f(\rho, \mu, \xi) = \rho + \frac{1}{\nu}\sum_{i=1}^{N} C_i \xi_i$

constraint convexity: $\begin{cases} g_{1,i}(\rho, \mu, \xi) = \|x_i - \mu\|^2 - \rho - \xi_i \le 0, \ \forall i \\ g_{2,i}(\rho, \mu, \xi) = -\xi_i \le 0, \\ g_{3,i}(\rho, \mu, \xi) = -\rho \le 0. \end{cases}$

slater's condition:

Existence of a strictly feasible point:

Choose any $\rho > 0$, $\xi_i > 0$, and $\mu \in \mathbb{R}^M$: $\|x_i - \mu\|^2 - \rho - \xi_i < 0, \forall i$

$\because$ $\xi_i > 0$ and $\rho > 0$, and μ can be adjusted appropriately, it is possible to satisfy the inequality strictly. $\Rightarrow$ Slater's condition is satisfied.

$\Rightarrow$ $\because$ the primal problem is convex and Slater's condition holds, strong duality applies.

The optimal values of the primal and dual problems are equal, and the duality gap is 0.

3. $\mathcal{L}(\rho, \mu, \xi, \alpha, \beta, \gamma) = \rho(1 - \sum_{i=1}^{N}\alpha_i - \gamma) + \sum_{i=1}^{N} \xi_i \left(\frac{C_i}{\nu} - \alpha_i - \beta_i\right) + \sum_{i=1}^{N} \alpha_i \|x_i - \mu\|^2$

Minimize ρ :

ρ is $1 - \sum_{i=1}^{N}\alpha_i - \gamma$ $\Rightarrow$ To prevent L from being unbound below $\Rightarrow$ $1 - \sum_{i=1}^{N}\alpha_i - \gamma = 0 \Rightarrow \gamma = 1 - \sum_{i=1}^{N}\alpha_i \xrightarrow{\gamma \ge 0 \text{ and } \alpha_i \ge 0} \sum_{i=1}^{N}\alpha_i \le 1$

Minimize ξ_i:

ξ_i is $\frac{C_i}{\nu} - \alpha_i - \beta_i$. To prevent $\mathcal{L}$ from being unbounded below ($\xi_i \ge 0$) $\Rightarrow \frac{C_i}{\nu} - \alpha_i - \beta_i = 0 \Rightarrow \beta_i = \frac{C_i}{\nu} - \alpha_i \xrightarrow{\beta_i \ge 0} \alpha_i \le \frac{C_i}{\nu}$.

Minimize μ:

$L_\mu = \sum_{i=1}^{N} \alpha_i \|x_i - \mu\|^2 \longrightarrow \nabla_\mu L = -2\sum_{i=1}^{N}\alpha_i(x_i - \mu) = 0 \Rightarrow \mu = \frac{\sum_{i=1}^{N}\alpha_i x_i}{\sum_{i=1}^{N}\alpha_i}$

Minimize Value of the Lagrangian:

$L_\mu = \sum_{i=1}^{N} \alpha_i \left\| x_i - \frac{\sum_{j=1}^{N}\alpha_j x_j}{\sum_{j=1}^{N}\alpha_j} \right\|^2 = \sum_{i=1}^{N}\alpha_i \|x_i\|^2 - \frac{\|\sum_{i=1}^{N}\alpha_i x_i\|^2}{\sum_{i=1}^{N}\alpha_i}$

$\theta(\alpha, \beta, \gamma) = \sum_{i=1}^{N}\alpha_i \|x_i\|^2 - \frac{\|\sum_{i=1}^{N}\alpha_i x_i\|^2}{\sum_{i=1}^{N}\alpha_i}$, $\begin{cases} \beta_i = \frac{C_i}{\nu} - \alpha_i \\ \gamma = 1 - \sum_{i=1}^{N}\alpha_i \end{cases}$

Constraints on Dual Variables:

$\alpha_i \ge 0$

$\beta_i \ge 0 \Rightarrow \alpha_i \le \frac{C_i}{\nu}$

$\gamma \ge 0 \Rightarrow \sum_{i=1}^{N}\alpha_i \le 1$

$\Rightarrow$ explicit expression: $\theta(\alpha, \beta, \gamma) = \sum_{i=1}^{N}\alpha_i \|x_i\|^2 - \frac{1}{\sum_{i=1}^{N}\alpha_i} \left\| \sum_{i=1}^{N}\alpha_i x_i \right\|^2$

with $\beta_i = \frac{C_i}{\nu} - \alpha_i$ and $\gamma = 1 - \sum_{i=1}^{N}\alpha_i$

4. Show that the dual problem can be simplified as

$$\begin{array}{ll} \text{maximize} & \|\alpha\|_1 \left(\sum_{i=1}^N \hat{\alpha}_i \|\mathbf{x}_i\|^2 - \sum_{1\le i,j\le N} \hat{\alpha}_i\hat{\alpha}_j \mathbf{x}_i^T \mathbf{x}_j\right) \\ \text{subject to} & \sum_{i=1}^N \alpha_i \le 1 \\ \text{variables} & 0 \le \alpha_i \le \frac{C_i}{\nu}, i \in [1, N] \end{array} \tag{18}$$

where $\|\alpha\|_1 = \sum_{i=1}^N \alpha_i$ and $\alpha_i = \|\alpha\|_1 \hat{\alpha}_i$.

4 let $\|\alpha\|_1 = \sum_{i=1}^N \alpha_i$, $\hat{\alpha}_i = \frac{\alpha_i}{\|\alpha\|_1}$ $\Rightarrow \alpha_i = \|\alpha\|_1 \hat{\alpha}_i$

$$\theta(\alpha) = \sum_{i=1}^N \alpha_i \|x_i\|^2 - \frac{1}{\sum_{i=1}^N \alpha_i}\left\|\sum_{i=1}^N \alpha_i x_i\right\|^2 \xrightarrow{\alpha_i = \|\alpha\|_1 \hat{\alpha}_i} \theta(\alpha) = \|\alpha\|_1 \sum_{i=1}^N \hat{\alpha}_i \|x_i\|^2 - \frac{1}{\|\alpha\|_1}\left\| \|\alpha\|_1 \sum_{i=1}^N \hat{\alpha}_i x_i\right\|^2 \Rightarrow \theta(\alpha) = \|\alpha\|_1 \left(\sum_{i=1}^N \hat{\alpha}_i \|x_i\|^2 - \left\|\sum_{i=1}^N \hat{\alpha}_i x_i\right\|^2\right)$$

$$\left\|\sum_{i=1}^N \hat{\alpha}_i x_i\right\|^2 = \left(\sum_{i=1}^N \hat{\alpha}_i x_i\right)^T \left(\sum_{j=1}^N \hat{\alpha}_j x_j\right) = \sum_{i=1}^N \sum_{j=1}^N \hat{\alpha}_i \hat{\alpha}_j x_i^T x_j$$

$$\theta(\alpha) = \|\alpha\|_1 \left(\sum_{i=1}^N \hat{\alpha}_i \|x_i\|^2 - \sum_{i=1}^N \sum_{j=1}^N \hat{\alpha}_i \hat{\alpha}_j x_i^T x_j\right)$$

Simplify the constraints:

$$\begin{array}{l} \alpha_i \ge 0 \Rightarrow \hat{\alpha}_i \ge 0 \\ \alpha_i \le \frac{C_i}{V} \Rightarrow \|\alpha\|_1 \hat{\alpha}_i \le \frac{C_i}{V} \\ \sum_{i=1}^N \alpha_i \le 1 \Rightarrow \|\alpha\|_1 \le 1 \end{array} \Rightarrow \begin{cases} \hat{\alpha}_i \ge 0 \\ \hat{\alpha}_i \le \frac{C_i}{V\|\alpha\|_1} \\ \|\alpha\|_1 \le 1 \end{cases}$$

$\because \|\alpha\|_1 \le 1$, $\frac{C_i}{V\|\alpha\|_1} \ge \frac{C_i}{V}$

To keep the upper bound tight, we use $\alpha_i \le \frac{C_i}{V}$.

objective Function:

Maximize $\theta(\alpha) = \|\alpha\|_1 \left(\sum_{i=1}^N \hat{\alpha}_i \|x_i\|^2 - \sum_{i=1}^N \sum_{j=1}^N \hat{\alpha}_i \hat{\alpha}_j x_i^T x_j\right)$

$$\begin{cases} \sum_{i=1}^N \alpha_i \le 1 \\ 0 \le \alpha_i \le \frac{C_i}{V}, \ \forall i \in [1, N] \end{cases} \Rightarrow \|\alpha\|_1 = \sum_{i=1}^N \alpha_i , \ \alpha_i = \|\alpha\|_1 \hat{\alpha}_i$$

5. Suppose $(\bar{\rho}, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\xi}})$ and $(\bar{\alpha}, \bar{\beta}, \bar{\gamma})$ are optimal solutions to problems (16) and (17), respectively.

   (a) Show that $\|\bar{\alpha}\|_1 \bar{\boldsymbol{\mu}} = \sum_{i=1}^N \bar{\alpha}_i \mathbf{x}_i$.

   (b) Show that

$$\bar{\rho} \in \arg min_{\rho \geq 0} \left(\rho + \frac{1}{\nu} \sum_{i=1}^N C_i \max(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \rho, 0) \right).$$

   (c) Show that

$$\min \left\{ \rho \geq 0 : \sum_{i:\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \rho} C_i \leq \nu \right\} \leq \bar{\rho} \leq \min \left\{ \rho \geq 0 : \sum_{i:\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \rho} C_i < \nu \right\}. \tag{19}$$

   (d) Prove that $\bar{\xi}_i = \max\left(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \bar{\rho}, 0\right)$ for each $i \in [1, N]$.

   (e) Prove that

$$\begin{cases} \bar{\alpha}_i = C_i/\nu & \text{, if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \bar{\rho} \\ \bar{\alpha}_i = 0 & \text{, if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 < \bar{\rho} \\ 0 \leq \bar{\alpha}_i \leq C_i/\nu & \text{, if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 = \bar{\rho} \end{cases}.$$

6. Suppose $C_i = 1/n$ for each $i \in [1, n]$. What is the physical meaning of ν?

5 (a) minimization over μ : $\mu = \dfrac{\sum_{i=1}^N \alpha_i x_i}{\sum_{i=1}^N \alpha_i}$

optimal point : $\bar{\mu} = \dfrac{\sum_{i=1}^N \bar{\alpha}_i x_i}{\sum_{i=1}^N \bar{\alpha}_i} = \dfrac{\sum_{i=1}^N \bar{\alpha}_i x_i}{\|\bar{\alpha}\|_1} \Rightarrow \|\bar{\alpha}\|_1 \bar{\mu} = \sum_{i=1}^N \bar{\alpha}_i x_i$

(b) Given $\bar{\mu}$, the primal problem reduces to minimizing over ρ and ξ : $\min_{\rho \geq 0, \xi \geq 0} \left(\rho + \frac{1}{\nu} \sum_{i=1}^N C_i \xi_i \right)$

$\|x_i - \bar{\mu}\|^2 - \rho - \xi_i \leq 0$, $\forall i$

optimal ξ_i : $\xi_i = \max\left(\|x_i - \bar{\mu}\|^2 - \rho, 0\right)$

objective function : $\min_{\rho \geq 0} \left(\rho + \frac{1}{\nu} \sum_{i=1}^N C_i \max(\|x_i - \bar{\mu}\|^2 - \rho, 0) \right) \Rightarrow \bar{\rho}$ minimizes this expression.

(c) $\bar{\rho}$ minimizes : $\phi(\rho) = \rho + \frac{1}{\nu} \sum_{i=1}^N C_i \max\left(\|x_i - \bar{\mu}\|^2 - \rho, 0\right)$

The derivative with respect to ρ is : $\phi'(\rho) = 1 - \frac{1}{\nu} \sum_{i:\|x_i - \bar{\mu}\|^2 > \rho} C_i$

Setting $\phi'(\bar{\rho}) = 0$: $1 - \frac{1}{\nu} \sum_{i:\|x_i - \bar{\mu}\|^2 > \bar{\rho}} C_i = 0 \Rightarrow \sum_{i:\|x_i - \bar{\mu}\|^2 > \bar{\rho}} C_i = \nu$

$\because \phi(\rho)$ is convex and piecewise linear with kinks at $\rho = \|x_i - \bar{\mu}\|^2$, $\bar{\rho}$ falls within the interval where the cumulative sum of C_i transitions from greater than ν to less than or equal to ν. $\therefore$ the inequality holds.

(d) complementarity Slackness : $\bar{\alpha}_i\left(\|x_i - \bar{\mu}\|^2 - \bar{\rho} - \bar{\xi}_i\right) = 0$

primal feasibility : $\|x_i - \bar{\mu}\|^2 - \bar{\rho} - \bar{\xi}_i \leq 0$

dual feasibility : $\bar{\alpha}_i \geq 0$

$\Rightarrow$ If $\|x_i - \bar{\mu}\|^2 - \bar{\rho} - \bar{\xi}_i < 0$, then $\bar{\alpha}_i = 0$

If $\bar{\xi}_i > 0$, then from the KKT conditions, $\bar{\beta}_i = 0$ and $\bar{\alpha}_i = \frac{C_i}{\nu}$

$\therefore \bar{\xi}_i = \max\left(\|x_i - \bar{\mu}\|^2 - \bar{\rho}, 0\right)$

(e) If $\|x_i - \bar{\mu}\|^2 > \bar{\rho}$:

$\bar{\xi}_i > 0$.

$\bar{\beta}_i \bar{\xi}_i = 0 \Rightarrow \bar{\beta}_i = 0$

$\bar{\beta}_i = \frac{C_i}{\nu} - \bar{\alpha}_i \Rightarrow \bar{\alpha}_i = \frac{C_i}{\nu}$

If $\|x_i - \bar{\mu}\|^2 < \bar{\rho}$:

$\bar{\xi}_i = 0$

primal feasibility : Constraint is strictly satisfied

Complementarity Slackness : $\bar{\alpha}_i = 0$

If $\|x_i - \bar{\mu}\|^2 = \bar{\rho}$:

$\bar{\xi}_i = 0$

Complementarity Slackness : $\bar{\alpha}_i \times 0 = 0$ (any $\bar{\alpha}_i \geq 0$)

Dual Feasibility : $0 \leq \bar{\alpha}_i \leq \frac{C_i}{\nu}$

6 $\sum_{i:\|x_i - \bar{\mu}\|^2 > \bar{\rho}} \frac{1}{n} = \nu \Rightarrow \dfrac{\text{number of outliers}}{n} = \nu$

$\therefore \nu$ represents the fraction of data points allowed to be outliers. (the proportion of points that can lie outside the minimal enclosing sphere)

* ν is the upper bound on the fraction of outliers in the data set. It controls the trade-off between the sphere's radius and the number of data points allowed outside the sphere. A smaller ν results in a tighter sphere encompassing more data points, while a larger ν allows for more outliers.