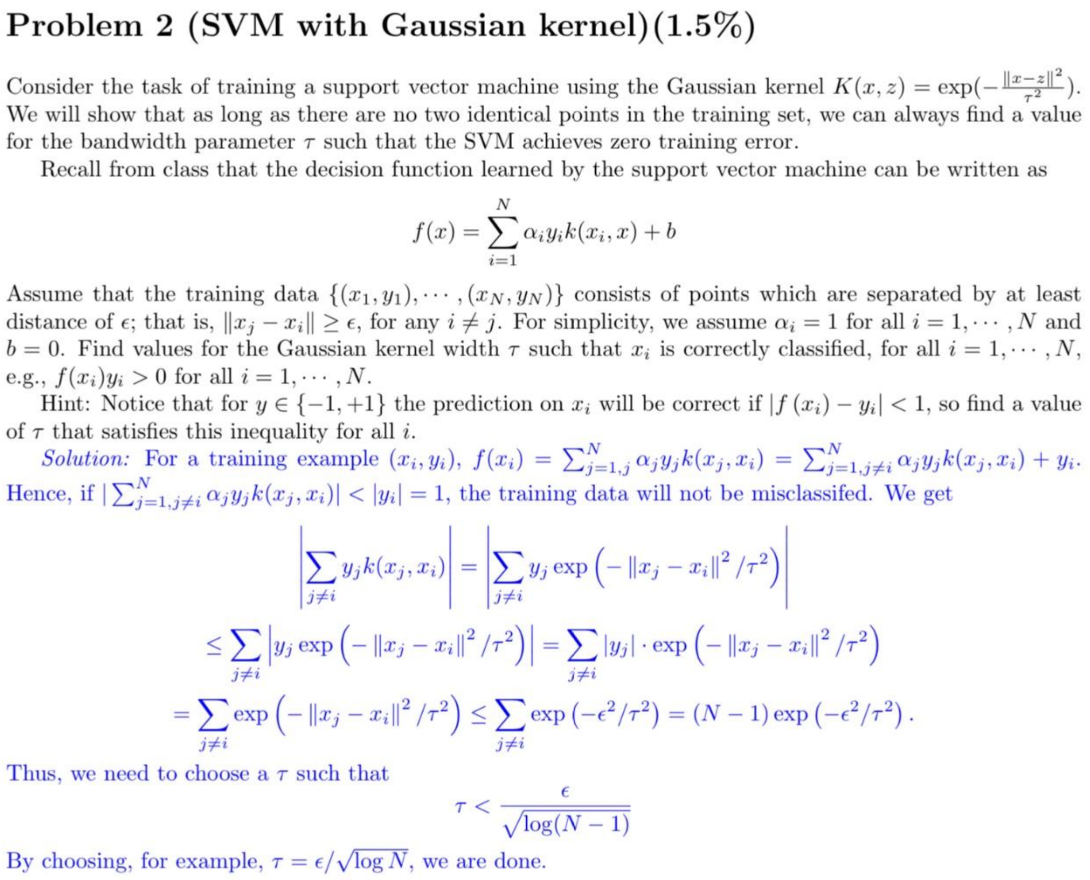
Problem 1

F

F

T



Problem 2 (SVM with Gaussian kernel)(1.5%)

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 $i = 1, \cdots, N$ 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.

Solution: For a training example (x_i, y_i), $f(x_i) = \sum_{j=1,j}^{N} \alpha_j y_j k(x_j, x_i) = \sum_{j=1,j\neq i}^{N} \alpha_j y_j k(x_j, x_i) + y_i$. Hence, if $|\sum_{j=1,j\neq i}^{N} \alpha_j y_j k(x_j, x_i)| < |y_i| = 1$, the training data will not be misclassifed. We get

$$\begin{aligned}
\left|\sum_{j\neq i} y_j k(x_j, x_i)\right| &= \left|\sum_{j\neq i} y_j \exp\left(-\left\|x_j - x_i\right\|^2/\tau^2\right)\right| \\
\leq \sum_{j\neq i}\left|y_j \exp\left(-\left\|x_j - x_i\right\|^2/\tau^2\right)\right| &= \sum_{j\neq i} |y_j| \cdot \exp\left(-\left\|x_j - x_i\right\|^2/\tau^2\right) \\
= \sum_{j\neq i} \exp\left(-\left\|x_j - x_i\right\|^2/\tau^2\right) &\leq \sum_{j\neq i} \exp\left(-\epsilon^2/\tau^2\right) = (N-1)\exp\left(-\epsilon^2/\tau^2\right).
\end{aligned}$$

Thus, we need to choose a τ such that

$$\tau < \frac{\epsilon}{\sqrt{\log(N-1)}}$$

By choosing, for example, $\tau = \epsilon/\sqrt{\log N}$, we are done.

Problem 3 (Support Vector Regression)(2%)

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

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \tag{1}$$

$$\text{s.t. } y_i - w^T x_i - b \leq \epsilon + \xi_i \qquad i = 1, \ldots, m \tag{2}$$

$$w^T x_i + b - y_i \leq \epsilon + \xi_i \qquad i = 1, \ldots, m \tag{3}$$

$$\xi_i \geq 0 \qquad i = 1, \ldots, m \tag{4}$$

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 (9), 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

(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 $e = y_i - \left(\bar{w}^T x_i + \bar{b}\right)$ Prove that

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

(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.

Solution:

1. Let $\alpha_i, \alpha_i^*, \beta \geq 0 (i = 1, \cdots, m)$ be the Lagrange multiplier for the primal problem. Then the Lagrangian can be written as:

$$\begin{aligned} & L\left(w, b, \xi, \alpha, \alpha^*, \beta, \right) \\ & = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \beta_i \xi_i \\ & - \sum_{i=1}^{m} \alpha_i \left(\epsilon + \xi_i - y_i + w^T x_i + b\right) \\ & - \sum_{i=1}^{m} \alpha_i^* \left(\epsilon + \xi_i + y_i - w^T x_i - b\right) \end{aligned} \tag{7}$$

2. Note that by $\alpha_i^{(*)}$, we refer to α_i and α_i^*. First, the dual function can be written as:

$$\theta(\alpha,\alpha^*,\beta) = \inf_{w,b,\xi} L\left(w,b,\xi,\alpha,\alpha^*,\beta\right) \tag{8}$$

Now, taking the derivatives of Lagrangian w.r.t. all primal variables, we get

$$\frac{\partial}{\partial w}L = w - \sum_{i=1}^m \left(\alpha_i - \alpha_i^*\right)x_i = 0 \Rightarrow w = \sum_{i=1}^m \left(\alpha_i - \alpha_i^*\right)x_i \tag{9}$$

$$\frac{\partial}{\partial b}L = \sum_{i=1}^m \left(\alpha_i^* - \alpha_i\right) = 0 \tag{10}$$

$$\frac{\partial}{\partial \xi}L = C - \alpha_i^{(*)} - \beta_i = 0 \tag{11}$$

Note that

$$\begin{aligned}\theta_D\left(\alpha,\alpha^*,\beta\right) = \frac{1}{2}\|w\|^2 - \epsilon\sum_{i=1}^m\left(\alpha_i+\alpha_i^*\right) + \sum_{i=1}^m y_i\left(\alpha_i-\alpha_i^*\right) + b\sum_{i=1}^m\left(\alpha_i^*-\alpha_i\right)\\ + \sum_{i=1}^m\left(\alpha_i^*-\alpha_i\right)w^Tx_i + \sum_{i=1}^m(C-\beta_i-\alpha_i-\alpha_i^*)\xi_i\end{aligned} \tag{12}$$

By the above equation(9)(10) and (11), we get

$$\begin{aligned}\theta_D\left(\alpha,\alpha^*\right) &= \frac{1}{2}\|w\|^2 - \epsilon\sum_{i=1}^m\left(\alpha_i+\alpha_i^*\right) + \sum_{i=1}^m y_i\left(\alpha_i-\alpha_i^*\right) + \sum_{i=1}^m\left(\alpha_i^*-\alpha_i\right)w^Tx_i\\ &= \frac{1}{2}\left\|\sum_{i=1}^m\left(\alpha_i-\alpha_i^*\right)x_i\right\|^2 - \sum_{i=1}^m\left(\alpha_i-\alpha_i^*\right)\left(\sum_{j=1}^m\left(\alpha_j-\alpha_j^*\right)x_j^Tx_i\right)\\ &- \epsilon\sum_{i=1}^m\left(\alpha_i+\alpha_i^*\right) + \sum_{i=1}^m y_i\left(\alpha_i-\alpha_i^*\right)\\ &= -\frac{1}{2}\sum_{i=1,j=1}^m\left(\alpha_i-\alpha_i^*\right)\left(\alpha_j-\alpha_j^*\right)x_i^Tx_j - \epsilon\sum_{i=1}^m\left(\alpha_i+\alpha_i^*\right) + \sum_{i=1}^m y_i\left(\alpha_i-\alpha_i^*\right)\end{aligned} \tag{13}$$

Now, the dual problem can be formulated as:

$$\begin{aligned}\max_{\alpha_i,\alpha_i^*} & -\frac{1}{2}\sum_{i=1,j=1}^m\left(\alpha_i-\alpha_i^*\right)\left(\alpha_j-\alpha_j^*\right)x_i^Tx_j - \epsilon\sum_{i=1}^m\left(\alpha_i+\alpha_i^*\right) + \sum_{i=1}^m y_i\left(\alpha_i-\alpha_i^*\right)\\ \text{s.t.} \quad & \sum_{i=1}^m\left(\alpha_i^*-\alpha_i\right) = 0\\ & 0 \le \alpha_i,\alpha_i^* \le C\end{aligned} \tag{14}$$

3. (a) Write the primal problem in the form:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \max\left(|y_i - \left(w^Tx_i+b\right)| - \epsilon, 0\right)$$

Since $\bar{w}$ is optimal, the optimal bias $\bar{b}$ is

$$\operatorname*{argmin}_b \sum_{i=1}^m \max\left(\left|y_i - \bar{w}^Tx_i + b\right| - \varepsilon, 0\right)$$

(b) Since $(\bar{w}, \bar{b}, \bar{\xi})$ and $(\bar{\alpha}, \bar{\alpha^*}, \bar{\beta})$ satisfies the KKT conditions that the following satisfies for all $i = 1, \cdots, N$:

$$\begin{array}{ll}
(S_1) \sum_{i=1}^{m} (\bar{\alpha_i} - \bar{\alpha_i}^*) = 0 & (P_1)\, y_i - \left(\bar{w}^T x_i + \bar{b}\right) - \epsilon - \bar{\xi_i} \leq 0 \\
(S_2)\, C = \bar{\alpha}_i + \bar{\alpha_i}^* + \bar{\beta_i} & (P_2) \left(\bar{w}^T x_i + \bar{b}\right) - y_i - \epsilon - \bar{\xi_i} \leq 0 \\
(S_3)\, \bar{w} = \sum_{i=1}^{m} (\bar{\alpha}_i - \bar{\alpha}_i^*)\, x_i & (P_3) - \bar{\xi_i} \leq 0 \\
(D_1)\, \bar{\alpha}_i, \bar{\alpha}_i^*, \bar{\beta}_i \geq 0 & (C_1)\, \bar{\alpha}_i \left(y_i - \left(\bar{w}^T x_i + \bar{b}\right) - \epsilon - \bar{\xi_i}\right) = 0 \\
 & (C_2)\, \bar{\alpha}_i^* \left(\left(\bar{w}^T x_i + \bar{b}\right) - y_i - \epsilon - \bar{\xi_i}\right) = 0 \\
 & (C_3)\, \bar{\beta}_i \left(-\bar{\xi_i}\right) = 0
\end{array}$$

(C_3) is rewritten as $(C - \bar{\alpha}_i - \bar{\alpha}_i^*)\xi_i = 0$ by (S_2)
Define $e = y_i - \left(\overline{w}^T x_i + \bar{b}\right)$

- If $|e| < \epsilon$, then $\bar{\alpha}_i = \bar{\alpha}_i^* = 0$ by $(C_1)(C_2)$, $\bar{\xi_i} = 0$ by (C_3)
- If $e = \epsilon$, then $\bar{\alpha}_i^* = 0$ by (C_2), $\bar{\xi_i} = 0$, $0 \leq \bar{\alpha}_i \leq C$ by $(C_1)(C_3)(D_1)$
- If $e = -\epsilon$, then $\bar{\alpha}_i = 0$ by (C_1), $\bar{\xi_i} = 0$, $0 \leq \bar{\alpha}_i^* \leq C$ by $(C_2)(C_3)(D_1)$
- If $e > \epsilon$, then $\bar{\alpha}_i^* = 0$ by (C_2), $\bar{\xi_i} \neq 0$ by (P_1), $\bar{\alpha}_i = C$ by (C_3), $\bar{\xi_i} = e - \epsilon$ by (C_1)
- If $e < -\epsilon$, then $\bar{\alpha}_i = 0$ by (C_1), $\bar{\xi_i} \neq 0$ by (P_2), $\bar{\alpha}_i^* = C$ by (C_3), $\bar{\xi_i} = -e - \epsilon$ by (C_2)

In fact, $\bar{\alpha_i}\bar{\alpha_i}^* = 0$ (its easily to prove by contradiction)

4. By equation (10) in (b), we have $w = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*)\, x_i$, then

$$f(w, x) = w^T x + b = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) x_i^T x + b = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) k(x_i, x) + b \tag{15}$$

This shows that the decision function can be written as a kernel form.

Problem 4

(35%) Hinge loss with L^1 regularization
Given data points $\mathbf{x}_1,\ldots,\mathbf{x}_N\in\mathbb{R}^m$ as well as their labels $y_1,\ldots,y_N\in\{\pm1\}$ and panelty coefficients $C_1,\ldots,C_N>0$, where each $\mathbf{x}_i=[x_{i,1},\ldots,x_{i,m}]^\mathsf{T}$ is a column vector, consider the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{w}\|_1+\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & \begin{array}{l} y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i+b)\geq1-\xi_i \\ \xi_i\geq0 \end{array} \quad ,\ i=1,\ldots,N \\ \text{variables} & \mathbf{w}\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^N \end{array} \tag{4}$$

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) **(5%)** Show that $(\bar{\mathbf{w}},\bar{b},\bar{\boldsymbol{\xi}})$ is an optimal solution of (4) 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:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=\sum_{j=1}^m(u_j+v_j)+\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & \begin{array}{ll} y_i((\mathbf{u}-\mathbf{v})^\mathsf{T}\mathbf{x}_i+b)\geq1-\xi_i & \\ \xi_i\geq0 & ,\ i=1,\ldots,N \\ u_j\geq0,\ v_j\geq0 & ,\ j=1,\ldots,m \end{array} \\ \text{variables} & \mathbf{u}\in\mathbb{R}^m,\mathbf{v}\in\mathbb{R}^m,b\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^N \end{array} \tag{5}$$

Following (a), we can now rewrite (5) as the following primal problem:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=\sum_{j=1}^m(u_j+v_j)+\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & \begin{array}{ll} g_i^1(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=1-\xi_i-y_i((\mathbf{u}-\mathbf{v})^\mathsf{T}\mathbf{x}_i+b)\leq0 & \\ g_i^2(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=-\xi_i\leq0 & ,\ i=1,\ldots,N \\ g_j^3(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=-u_j\leq0 & \\ g_j^4(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=-v_j\leq0 & ,\ j=1,\ldots,m \end{array} \\ \text{variables} & \mathbf{u}\in\mathbb{R}^m,\mathbf{v}\in\mathbb{R}^m,b\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^N \end{array} \tag{6}$$

as well as its Lagrangian dual problem:

$$\begin{array}{ll} \text{maximize} & \theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu})=\inf\{L(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi},\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu}):\mathbf{u}\in\mathbb{R}^m,\mathbf{v}\in\mathbb{R}^m,b\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^N\} \\ \text{subject to} & \begin{array}{ll} \alpha_i\geq0,\ \beta_i\geq0 & ,\ i=1,\ldots,N \\ \mu_j\geq0,\ \nu_j\geq0 & ,\ j=1,\ldots,m \end{array} \\ \text{variables} & \alpha\in\mathbb{R}^N,\beta\in\mathbb{R}^N,\boldsymbol{\mu}\in\mathbb{R}^m,\boldsymbol{\nu}\in\mathbb{R}^m \end{array} \tag{7}$$

where L denotes the Lagranian function.

(b) **(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:

$$L(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi},\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu})$$

$$=\mathbf{1}^\mathsf{T}(\mathbf{u}+\mathbf{v})+\sum_{i=1}^N C_i\xi_i+\sum_{i=1}^N\alpha_i\Big(1-\xi_i-y_i((\mathbf{u}-\mathbf{v})^\mathsf{T}\mathbf{x}_i+b)\Big)+\sum_{i=1}^N\beta_i(-\xi_i)-\boldsymbol{\mu}^\mathsf{T}\mathbf{u}-\boldsymbol{\nu}^\mathsf{T}\mathbf{v}$$

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

(c) **(3%)** Show that (6) satisfies the Slater's condition.

(d) **(6%)** Show that

  (i) **(3%)** $\theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu})=-\infty$ unless the following conditions hold:

$$\sum_{i=1}^{N}\alpha_i y_i=0,\ \ \boldsymbol{\mu}=\mathbf{1}-\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i,\ \ \boldsymbol{\nu}=\mathbf{1}+\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i \tag{8a}$$

$$\alpha_i+\beta_i=C_i\ \ \forall i=1,\ldots,N \tag{8b}$$

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

  (ii) **(3%)** The stationary condition holds if and only if (8) is satisfied.

(e) **(2%)** Show that the dual problem (7) can be simplified as

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{N}\alpha_i \\ \text{subject to} & \sum_{i=1}^{N}\alpha_i y_i=0 \\ & -\mathbf{1}\leq\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i\leq\mathbf{1} \\ \text{variables} & 0\leq\alpha_i\leq C_i,\ \ i=1,\ldots,N \end{array} \tag{9}$$

(f) **(4%)** Write down the KKT conditions for primal/dual problems (6)(7).

(g) **(10%)** Suppose $(\bar{\mathbf{u}},\bar{\mathbf{v}},\bar{b},\bar{\boldsymbol{\xi}})$ and $(\bar{\alpha},\bar{\beta},\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\nu}})$ are optimal solutions to (6) and (7) respectively. Denote $\bar{\mathbf{w}}=\bar{\mathbf{u}}-\bar{\mathbf{v}}$. Show that

  (i) **(3%)** $\bar{u}_j=\max(\bar{w}_j,0)$ and $\bar{v}_j=\max(-\bar{w}_j,0)$. (Hint: Consider the cases $\sum_{i=1}^{N}\bar{\alpha}_i y_i x_{i,j}<1$ and $\sum_{i=1}^{N}\bar{\alpha}_i y_i x_{i,j}>-1$.)

  (ii) **(3%)** $\bar{w}_j=0$ unless $\sum_{i=1}^{N}\bar{\alpha}_i y_i x_{i,j}\in\{\pm 1\}$.

  (iii) **(4%)**

$$\begin{cases} \bar{\alpha}_i=C_i, & \bar{\xi}_i=1-y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+\bar{b}), & \text{if } y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+\bar{b})<1 \\ \bar{\alpha}_i=0, & \bar{\xi}_i=0, & \text{if } y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+\bar{b})>1 \\ 0\leq\bar{\alpha}_i\leq C_i, & \bar{\xi}_i=0, & \text{if } y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+\bar{b})=1 \end{cases}$$

(h) **(3%)** Is it true that $\bar{\mathbf{w}}$ must be a linear combination of $\mathbf{x}_1,\ldots,\mathbf{x}_N$? Justify your answers.

Solution:

(a) Denote $f_0(\mathbf{w},b,\boldsymbol{\xi})=\|\mathbf{w}\|_1+\sum_{i=1}^{N}C_i\xi_i$ as the objective function for (4). For each feasible solution $(\mathbf{w},b,\boldsymbol{\xi})$ to (4), one can construct $\mathbf{w}^+$ and $\mathbf{w}^-$ where $w_j^+=\max(w_j,0)$ and $w_j^-=\max(-w_j,0)$. Note that $(\mathbf{w}^+,\mathbf{w}^-,b,\boldsymbol{\xi})$ is a feasible solution to (5), and that

$$f_0(\mathbf{w},b,\boldsymbol{\xi})=f(\mathbf{w}^+,\mathbf{w}^-,b,\boldsymbol{\xi})\geq\text{Minimum of (5)}.$$

On the other hand, for each feasible solution $(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})$ to (5), note that $(\mathbf{u}-\mathbf{v},b,\boldsymbol{\xi})$ is a feasible solution to (4), and that

$$f(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})\geq f_0(\mathbf{u}-\mathbf{v},b,\boldsymbol{\xi})\geq\text{Minimum of (4)}.$$

Hence (4) and (5) have the same minimum value. As such, if $(\mathbf{w},b,\boldsymbol{\xi})$ is an optimal solution to (4), then $\mathbf{w}=\mathbf{w}^{+}-\mathbf{w}^{-}$ where $(\mathbf{w}^{+},\mathbf{w}^{-},b,\boldsymbol{\xi})$ is an optimal solution to (5); Conversely, if $(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})$ is an optimal solution to (5), then $(\mathbf{u}-\mathbf{v},b,\boldsymbol{\xi})$ is an optimal solution to (4).

(b)

$$\begin{aligned}
&L(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi},\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu})\\
&=f(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})+\sum_{i=1}^{N}\alpha_i g_i^1(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})+\sum_{i=1}^{N}\beta_i g_i^2(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})+\sum_{j=1}^{m}\mu_j g_j^3(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})+\sum_{j=1}^{m}\nu_j g_j^4(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})\\
&=\sum_{j=1}^{m}(u_j+v_j)+\sum_{i=1}^{N}C_i\xi_i+\sum_{i=1}^{N}\alpha_i\Big(1-\xi_i-y_i((\mathbf{u}-\mathbf{v})^{\mathsf{T}}\mathbf{x}_i+b)\Big)+\sum_{i=1}^{N}\beta_i(-\xi_i)+\sum_{j=1}^{m}\mu_j(-u_j)+\sum_{j=1}^{m}\nu_j(-v_j)\\
&=\mathbf{1}^{\mathsf{T}}(\mathbf{u}+\mathbf{v})+\sum_{i=1}^{N}C_i\xi_i+\sum_{i=1}^{N}\alpha_i\Big(1-\xi_i-y_i((\mathbf{u}-\mathbf{v})^{\mathsf{T}}\mathbf{x}_i+b)\Big)+\sum_{i=1}^{N}\beta_i(-\xi_i)-\boldsymbol{\mu}^{\mathsf{T}}\mathbf{u}-\boldsymbol{\nu}^{\mathsf{T}}\mathbf{v}
\end{aligned}$$

(c) (Omitted)

(d) Taking partial derivatives of the Lagranian function L over $\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}$ yields

$$\nabla_{\mathbf{u}}L=\mathbf{1}-\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i-\boldsymbol{\mu},\quad \nabla_{\mathbf{v}}L=\mathbf{1}+\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i-\boldsymbol{\nu}$$

$$\frac{\partial}{\partial b}L=-\sum_{i=1}^{N}\alpha_i y_i,\quad \frac{\partial}{\partial \xi_i}L=C_i-\alpha_i-\beta_i.$$

If (8) is satisfied, then $\theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu})=L(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi},\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu})=\sum_{i=1}^{N}\alpha_i$; Otherwise, as $\nabla_{\mathbf{u}}L,\nabla_{\mathbf{v}}L,\frac{\partial}{\partial b}L,\frac{\partial}{\partial \xi_i}L$ are constant (vectors or scalars) not identically zero, one has $\theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu})=-\infty$. In particular, the Stationary Condition, namely $\theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu})=L(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi},\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu})$, holds if and only if (8) is satisfied.

(e) From (d), it suffices to consider dual feasible solutions that satisfy (8), which leads to

$$\begin{array}{ll}
\text{maximize} & \sum_{i=1}^{N}\alpha_i\\
& \alpha_i\geq 0,\ \beta_i\geq 0,\ \alpha_i+\beta_i=C_i\\
\text{subject to} & \mu_j\geq 0,\ \nu_j\geq 0\\
& \sum_{i=1}^{N}\alpha_i y_i=0,\ \boldsymbol{\mu}=\mathbf{1}-\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i,\ \boldsymbol{\nu}=\mathbf{1}+\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i\\
\text{variables} & \alpha\in\mathbb{R}^N,\beta\in\mathbb{R}^N,\boldsymbol{\mu}\in\mathbb{R}^m,\ \boldsymbol{\nu}\in\mathbb{R}^m
\end{array}$$

which can be further simplified to (9).

(f) The **S**tationary, **P**rimal and **D**ual feasibility, and **C**omplementary slackness conditions are given by

$$\begin{array}{llll}
\text{(S1)} & \alpha_i+\beta_i=C_i & \text{(C1)} & \alpha_i g_i^1(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=\alpha_i\big(1-\xi_i-y_i((\mathbf{u}-\mathbf{v})^{\mathsf{T}}\mathbf{x}_i+b)\big)=0\\
\text{(S2)} & \sum_{i=1}^{N}\alpha_i y_i=0 & \text{(C2)} & \beta_i g_i^2(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=\beta_i(-\xi_i)=0\\
\text{(S3)} & \boldsymbol{\mu}=\mathbf{1}-\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i & \text{(C3)} & \mu_j g_j^3(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=\mu_j(-u_j)=0\\
\text{(S4)} & \boldsymbol{\nu}=\mathbf{1}+\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i & \text{(C4)} & \nu_j g_j^4(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=\nu_j(-v_j)=0
\end{array}$$

(P1)	$g_i^1(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=1-\xi_i-y_i((\mathbf{u}-\mathbf{v})^\mathsf{T}\mathbf{x}_i+b)\leq 0$	(D1)	$\alpha_i\geq 0$
(P2)	$g_i^2(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=-\xi_i\leq 0$	(D2)	$\beta_i\geq 0$
(P3)	$g_j^3(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=-u_j\leq 0$	(D3)	$\mu_j\geq 0$
(P4)	$g_j^4(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})=-v_j\leq 0$	(D4)	$\nu_j\geq 0$

(g) By (c), the strong duality theorem implies zero duality gap, so the primal and dual optimal solutions $(\bar{\mathbf{u}},\bar{\mathbf{v}},\bar{b},\bar{\boldsymbol{\xi}})$ and $(\bar{\alpha},\bar{\beta},\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\nu}})$ satisfy KKT conditions as given by (f). Note that

- If $\sum_{i=1}^N \bar{\alpha}_i y_i x_{i,j}<1$, then $\bar{\mu}_j>0$ by (S3), $\bar{u}_j=0$ by (C3), $\bar{v}_j\geq 0$ by (P4).
- If $\sum_{i=1}^N \bar{\alpha}_i y_i x_{i,j}>-1$, then $\bar{\nu}_j>0$ by (S4), $\bar{v}_j=0$ by (C4), $\bar{u}_j\geq 0$ by (P3).

One concludes that $\bar{u}_j=\max(\bar{w}_j,0)$ and $\bar{v}_j=\max(-\bar{w}_j,0)$, and that $\bar{w}_j=0$ whenever $|\sum_{i=1}^N \bar{\alpha}_i y_i x_{i,j}|<1$. Since $|\sum_{i=1}^N \bar{\alpha}_i y_i x_{i,j}|\leq 1$ by (S3,S4,D3,D4), so $\bar{w}_j=0$ unless $\sum_{i=1}^N \bar{\alpha}_i y_i x_{i,j}\in\{\pm 1\}$. Finally,

- If $y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+b)<1$, then $\bar{\xi}_i>0$ by (P1), $\bar{\beta}_i=0$ by (C2), $\bar{\alpha}_i=C_i$ by (S1), $\bar{\xi}_i=1-y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+\bar{b})$ by (C1).
- If $y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+b)>1$, then $\bar{\alpha}_i=0$ by (P2,C1), $\bar{\beta}_i=C_i$ by (S1), $\bar{\xi}_i=0$ by (C2).
- If $y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+b)=1$, then $\bar{\xi}_i=0$ by (C1,C2,S1), $0\leq\bar{\alpha}_i\leq C_i$ by (S1,D1,D2).

(h) No. The representation theorem does not apply to the L^1-regularization scenario.

Problem 5 (Spherical one class SVM)(2%)(Bonus)

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)

$$\begin{array}{ll} \text{minimize} & R^2 + \frac{1}{\nu}\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & \left.\begin{array}{l} \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \leq R^2 + \xi_i \\ \xi_i \geq 0 \end{array}\right\} \forall i \in 1, N \\ & R \geq 0 \\ \text{variables} & R \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} = (\xi_1, ..., \xi_N) \in \mathbb{R}^N \end{array} \tag{26}$$

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 (26) in the form of primal problem:

$$\begin{array}{ll} \text{minimize} & f(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = \rho + \frac{1}{\nu}\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & \left.\begin{array}{l} g_{1,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 - \rho - \xi_i \leq 0 \\ g_{2,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = -\xi_i \leq 0 \end{array}\right\} \forall i \in 1, N \\ & g_3(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = -\rho \leq 0 \\ \text{variables} & \rho \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} \in \mathbb{R}^N \end{array} \tag{27}$$

as well its Lagrangian dual problem:

$$\begin{array}{ll} \text{maximize} & \theta(\alpha, \beta, \gamma) = \inf_{\rho\in\mathbb{R}, \boldsymbol{\mu}\in\mathbb{R}^M, \boldsymbol{\xi}\in\mathbb{R}^N} L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma) \\ \text{subject to} & \alpha_i \geq 0, \beta_i \geq 0 \ \forall i \in 1, N \\ & \gamma \geq 0 \\ \text{variables} & \alpha = (\alpha_1, ..., \alpha_N) \in \mathbb{R}^N, \beta = (\beta_1, ..., \beta_N) \in \mathbb{R}^N, \gamma \in \mathbb{R} \end{array} \tag{28}$$

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 (27) and (28) is zero.
3. Derive $\theta(\alpha, \beta, \gamma)$ in explicit form of dual variables α, β, γ.
4. Show that the dual problem can be simplified as

$$\begin{array}{ll} \text{maximize} & \|\alpha\|_1 \left(\sum_{i=1}^N \hat{\alpha}_i\|\mathbf{x}_i\|^2 - \sum_{1\leq i,j\leq N} \hat{\alpha}_i\hat{\alpha}_j\mathbf{x}_i^T\mathbf{x}_j\right) \\ \text{subject to} & \sum_{i=1}^N \alpha_i \leq 1 \\ \text{variables} & 0 \leq \alpha_i \leq \frac{C_i}{\nu}, i \in 1, N \end{array} \tag{29}$$

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

5. Suppose $(\bar{\rho}, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\xi}})$ and $(\bar{\alpha}, \bar{\beta}, \bar{\gamma})$ are optimal solutions to problems (27) and (28), respectively.
   (a) Show that $\|\bar{\alpha}\|_1\bar{\boldsymbol{\mu}} = \sum_{i=1}^N \bar{\alpha}_i\mathbf{x}_i$.
   (b) Show that
   $$\bar{\rho} \in_{\rho\geq 0} \left(\rho + \frac{1}{\nu}\sum_{i=1}^N C_i \max(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \rho, 0)\right).$$
   (c) Show that
   $$\min\left\{\rho \geq 0 : \sum_{i:\|\mathbf{x}_i-\bar{\boldsymbol{\mu}}\|^2>\rho} C_i \leq \nu\right\} \leq \bar{\rho} \leq \min\left\{\rho \geq 0 : \sum_{i:\|\mathbf{x}_i-\bar{\boldsymbol{\mu}}\|^2>\rho} C_i < \nu\right\}. \tag{30}$$
   (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?

Solution:

1. The Lagrangian function can be explicitly written as

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

2. We can verify that the problem satisfy the condition of strong duality theorem. Hence, the duality gap is zero.

3. Take partial deriviates of the Lagranian function L over $\rho, \boldsymbol{\mu}, \boldsymbol{\xi}$, one has

$$\frac{\partial}{\partial \rho} L = 1 - \sum_{i=1}^{N} \alpha_i - \gamma, \quad \nabla_{\boldsymbol{\mu}} L = 2 \sum_{i=1}^{N} \alpha_i (\boldsymbol{\mu} - \mathbf{x}_i), \quad \frac{\partial}{\partial \xi_i} L = \frac{C_i}{\nu} - \alpha_i - \beta_i.$$

   - If the following conditions hold:

     $$\sum_{i=1}^{N} \alpha_i + \gamma = 1, \ \alpha_i + \beta_i = \frac{C_i}{\nu} \ \forall i \in 1, N, \tag{32}$$

     then $\theta(\alpha, \beta, \gamma) = L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma)$ iff

     $$\|\alpha\|_1 \boldsymbol{\mu} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i, \tag{33}$$

     at which

     $$\theta(\alpha, \beta, \gamma) = \|\alpha\|_1 \left(\sum_{i=1}^{N} \hat{\alpha}_i \|\mathbf{x}_i\|^2 - \sum_{1 \le i,j \le N} \hat{\alpha}_i \hat{\alpha}_j \mathbf{x}_i^T \mathbf{x}_j \right) \tag{34}$$

   - Otherwise, as $\frac{\partial}{\partial \rho} L$, $\frac{\partial}{\partial \xi_1} L$, ..., $\frac{\partial}{\partial \xi_N} L$ are constants not identically zero, one has $\theta(\alpha, \beta, \gamma) = -\infty$.

4. By (c), one may rewrite (28) as

$$\begin{array}{ll} \text{maximize} & \|\alpha\|_1 \left(\sum_{i=1}^{N} \hat{\alpha}_i \|\mathbf{x}_i\|^2 - \sum_{1 \le i,j \le N} \hat{\alpha}_i \hat{\alpha}_j \mathbf{x}_i^T \mathbf{x}_j \right) \\ \text{subject to} & \left. \begin{array}{l} \alpha_i \ge 0, \beta_i \ge 0 \\ \alpha_i + \beta_i = \frac{C_i}{\nu} \end{array} \right\} \ \forall i \in 1, N \\ & \sum_{i=1}^{N} \alpha_i + \gamma = 1 \\ & \gamma \ge 0 \\ \text{variables} & \alpha \in \mathbb{R}^N, \beta \in \mathbb{R}^N, \gamma \in \mathbb{R} \end{array}$$

   which can be further simplified as (29).

5. By (b), the duality gap is zero, so the primal and dual optimal solutions satisfy the KKT conditions. Note that in (c) we have shown that the stationary condition $\theta(\alpha, \beta, \gamma) = L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma)$ holds iff (32) and (33) are both satisfied. As such, we may write down the KKT conditions, including **S**tationary, **P**rimal feasible, **D**ual feasible, and **C**omplimentary conditions, as follows (for all $i \in 1, N$)

$$\begin{array}{llll} \text{(S1)} & \bar{\alpha}_i + \bar{\beta}_i = C_i/\nu & \text{(D1)} & \bar{\alpha}_i \ge 0 \\ \text{(S2)} & \sum_{i=1}^{N} \bar{\alpha}_i + \bar{\gamma} = 1 & \text{(D2)} & \bar{\beta}_i \ge 0 \\ \text{(S3)} & \|\bar{\alpha}\|_1 \bar{\boldsymbol{\mu}} = \sum_{i=1}^{N} \bar{\alpha}_i \mathbf{x}_i & \text{(D3)} & \bar{\gamma} \ge 0 \\ \text{(P1)} & \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 \le \bar{\rho} + \bar{\xi}_i & \text{(C1)} & \bar{\alpha}_i \left(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \bar{\rho} - \bar{\xi}_i \right) = 0 \\ \text{(P2)} & \bar{\xi}_i \ge 0 & \text{(C2)} & \bar{\beta}_i \bar{\xi}_i = 0 \\ \text{(P3)} & \bar{\rho} \ge 0 & \text{(C3)} & \bar{\gamma} \bar{\rho} = 0 \end{array}$$

   Note that (27) can be rewritten as the following optimization problem

$$\begin{array}{ll} \text{minimize} & \rho + \frac{1}{\nu} \sum_{i=1}^{N} C_i \max(\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 - \rho, 0) \\ \text{subject to} & \rho \ge 0 \\ \text{variables} & \rho \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M \end{array}$$

   Given the optimal centroid $\bar{\boldsymbol{\mu}}$, the optimal squared radius $\bar{\rho}$ is therefore given by

$$\bar{\rho} \in_{\rho \ge 0} \underbrace{\left(\rho + \frac{1}{\nu} \sum_{i=1}^{N} C_i \max(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \rho, 0) \right)}_{G(\rho)}.$$

Note that G is a convex and continuous function with right derivative

$$\partial_+ G(\rho) := \lim_{\Delta\rho \downarrow 0} \frac{G(\rho + \Delta\rho) - G(\rho)}{\Delta\rho} = 1 - \frac{F(\rho)}{\nu},$$

where $F : \rho \in [0, \infty) \mapsto \sum_{i:\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \rho} C_i$ is a non-increasing function. Let $\rho_1 = \min\{\rho \geq 0 : F(\rho) \leq \nu\}$ and $\rho_2 = \min\{\rho \geq 0 : F(\rho) < \nu\}$.

- If $\rho < \rho_1$, then $F(\rho) > \nu$ and $\partial_+ G(\rho) < 0$. Thus G is strictly decreasing on $[0, \rho_1]$.
- If $\rho \geq \rho_2$, then $F(\rho) < \nu$ and $\partial_+ G(\rho) > 0$. Thus G is strictly increasing on $[\rho_2, \infty)$.
- If $\rho \in [\rho_1, \rho_2)$, then $F(\rho) = \nu$ and $\partial_+ G(\rho) = 0$. Thus G remains constant on $[\rho_1, \rho_2]$.

- If $\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \bar{\rho}$, then $\bar{\xi}_i > 0$ by (P1), $\bar{\beta}_i = 0$ by (C2), $\bar{\alpha}_i = C_i/\nu$ by (S1), $\bar{\xi}_i = \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \bar{\rho}$ by (C1).
- If $\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 < \bar{\rho}$, then $\bar{\alpha}_i = 0$ by (P2,C1), $\bar{\beta}_i = C_i/\nu$ by (S1), $\bar{\xi}_i = 0$ by (C2).
- If $\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 = \bar{\rho}$, then $\bar{\xi}_i = 0$ by (S1,C1,C2), and $0 \leq \bar{\alpha}_i \leq C_i/\nu$ by (D1,D2,S1).

6. If $C_i = \frac{1}{n}$, by 5(c), ν is the acceptable violation rate of the data points. For example, if $\nu = 0.1$, $\bar{\rho}$ is chosen to be the value that 10% of the data points are outside of the ball.