ML2024 Fall Assignment 4

Wei-Chen Chang, R12227118

Due: Nov, 22 2024

Problem 1

1. T
2. F
3. T

Problem 2

It suffice to show that $|f(x_i) - y_i| < 1 \quad \forall i = 1, .., N$. Hence:

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

Let N^+, N^- be the number of cases where $y_i = +1/y_i = -1$ respectively, the inequality becomes:

$$\begin{aligned}
|f(x_i) - y_i| &\leq \left| \sum_{j: y_j = 1, j \neq i} \exp\left(-\frac{\epsilon^2}{\tau^2}\right) - \sum_{j: y_j = -1, j \neq i} \exp\left(-\frac{\epsilon^2}{\tau^2}\right) \right| \\
&= \begin{cases} |N^+ - 1 - N^-| \exp\left(-\frac{\epsilon^2}{\tau^2}\right), & y_i = +1 \\ |N^+ - (N^- - 1)| \exp\left(-\frac{\epsilon^2}{\tau^2}\right), & y_i = -1 \end{cases} \\
&\leq (|N^+ - N^-| + 1) \exp\left(-\frac{\epsilon^2}{\tau^2}\right) < 1.
\end{aligned}$$

Solve the inequality, we have:

$$\tau < \frac{\epsilon}{\sqrt{\log\left(|N^+ - N^-| + 1\right)}}$$

Problem 3

(a)

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

(b)

Let $\theta(\alpha,\alpha^*,\beta) = \inf_{\tilde{w},\tilde{b},\tilde{\xi}} \mathcal{L}(\tilde{w},\tilde{b},\tilde{\xi},\alpha,\alpha^*,\beta)$. The dual problem can be formulated as:

$$\begin{aligned} \text{maximize} \quad & \theta(\alpha,\alpha^*,\beta) \\ \text{subject to} \quad & \alpha \geq 0, \alpha^* \geq 0, \beta \geq 0 \quad i = 1,...,m \\ \text{variables} \quad & \alpha_i \in \mathbb{R}, \alpha_i^* \in \mathbb{R}, \beta_i \in \mathbb{R} \quad i = 1,...,m. \end{aligned}$$

Moreover, we can simplify $\theta(\alpha,\alpha^*,\beta)$ to minimize over Lagrangian. First we take the derivatives of Lagrangian w.r.t. w, b, ξ:

$$\nabla_w \mathcal{L} = w - \sum_{i=1}^m (\alpha_i - \alpha_i^*)x_i \tag{1}$$

$$\frac{\partial}{\partial b}\mathcal{L} = \sum_{i=1}^m (\alpha_i^* - \alpha_i) \tag{2}$$

$$\frac{\partial}{\partial \xi_i}\mathcal{L} = C - \alpha_i - \alpha_i^* - \beta_i, \quad \forall i = 1,...,m. \tag{3}$$

Note that if (2), (3) $= 0$, then (1) implies that $w = \sum_{i=1}^m (\alpha_i - \alpha_i^*)x_i$ to minimize $\theta(\alpha,\alpha^*,\beta)$. Otherwise, if one of the (2), (3) $\neq 0$, $\mathcal{L} = -\infty$. Hence we can further simplify $\theta(\alpha,\alpha^*,\beta)$ as:

$$\theta(\alpha,\alpha^*,\beta) = \sum_{i=1}^m (\alpha_i - \alpha_i^*)y_i - \frac{1}{2}\sum_{0\leq i,j\leq m} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^T x_j - \epsilon\sum_{i=1}^m (\alpha_i + \alpha_i^*),$$

and dual problem as:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^m (\alpha_i - \alpha_i^*)y_i - \frac{1}{2}\sum_{0\leq i,j\leq m} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^T x_j - \epsilon\sum_{i=1}^m (\alpha_i + \alpha_i^*) \\ \text{subject to} \quad & \alpha \geq 0, \alpha^* \geq 0, \beta \geq 0,\ \alpha_i + \alpha_i^* + \beta_i = C, \quad i = 1,...,m \\ & \sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0, \\ \text{variables} \quad & \alpha_i \in \mathbb{R}, \alpha_i^* \in \mathbb{R}, \beta_i \in \mathbb{R} \quad i = 1,...,m. \end{aligned}$$

(c)

1. Since $\bar{b}, \bar{w}, \bar{\xi}$ are primal optimal, then they are primal feasible, we have:

$$
\begin{aligned}
y_i - (\bar{w}^T x_i + \bar{b}) &\le \epsilon + \bar{\xi}_i \\
(\bar{w}^T x_i + \bar{b}) - y_i &\le \epsilon + \bar{\xi}_i \\
\bar{\xi}_i &\ge 0.
\end{aligned}
$$

   Combining these equations, we get $\bar{\xi}_i \ge \max\{|y_i - (\bar{w}^T x_i + \bar{b})| - \epsilon, 0\}$.

   Recall the target function in primal problem: $\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \xi_i$, replace ξ_i by $\max\{|y_i - (\bar{w}^T x_i + \bar{b})| - \epsilon, 0\}$ then didn't change its solution. Also, since $\bar{b}$ is the minimizer and the optimal $w = \bar{w}$ is given, we have:

$$
\bar{b} = \arg\min_b C \sum_{i=1}^m \max\left\{|y_i - (\bar{w}^T x_i + b)| - \epsilon, 0\right\}
$$

2. Since Slater's condition holds, the problem has zero-gap. $(\bar{b}, \bar{w}, \bar{\xi})$ and $(\bar{\alpha}, \bar{\alpha}^*, \bar{\beta})$ are optimal to primal and dual problem respectively implies that KKT conditions hold. They are:

Stationary

$$
\begin{aligned}
&\text{(S1)} \quad \bar{w} = \sum_{i=1}^m (\bar{\alpha}_i - \bar{\alpha}_i^*) x_i, \\
&\text{(S2)} \quad \sum_{i=1}^m (\bar{\alpha}_i - \bar{\alpha}_i^*) = 0, \\
&\text{(S3)} \quad \bar{\alpha}_i + \bar{\alpha}_i^* + \bar{\beta}_i = C, && \forall i = 1, \dots, m
\end{aligned}
$$

Complementary Slackness

$$
\begin{aligned}
&\text{(C1)} \quad \bar{\alpha}_i (y_i - \bar{w}^T x_i - b - \bar{\xi}_i - \epsilon) = 0, && \forall i = 1, \dots, m \\
&\text{(C2)} \quad \bar{\alpha}_i^* (\bar{w}^T x_i + \bar{b} - y_i - \bar{\xi}_i - \epsilon) = 0, && \forall i = 1, \dots, m \\
&\text{(C3)} \quad \bar{\beta}_i \bar{\xi}_i = 0, && \forall i = 1, \dots, m
\end{aligned}
$$

Feasibility

$$
\begin{aligned}
&\text{(P1)} \quad y_i - \bar{w}^T x_i - \bar{b} \le \epsilon + \bar{\xi}_i, && \forall i = 1, \dots, m \\
&\text{(P2)} \quad \bar{w}^T x_i + \bar{b} - y_i \le \epsilon + \bar{\xi}_i, && \forall i = 1, \dots, m \\
&\text{(P3)} \quad \bar{\xi}_i \ge 0, && \forall i = 1, \dots, m \\
&\text{(D1)} \quad \bar{\alpha}_i \ge 0, && \forall i = 1, \dots, m \\
&\text{(D2)} \quad \bar{\alpha}_i^* \ge 0, && \forall i = 1, \dots, m \\
&\text{(D3)} \quad \bar{\beta}_i \ge 0, && \forall i = 1, \dots, m.
\end{aligned}
$$

Then one can discuss it case by case:

1. If $|e| < \epsilon$, then $|e| - \epsilon < 0 \implies \begin{cases} e - \epsilon \leq |e| - \epsilon < 0 \\ -e - \epsilon \leq |e| - \epsilon < 0 \end{cases}$, then $\bar{\alpha}_i = 0$ by (D1, C1); and $\bar{\alpha}_i^* = 0$ by (D2, C2), then $\bar{\beta}_i = C > 0$ by (S3), lastly $\bar{\xi}_i = 0$ by (C3).

2. If $e = \epsilon$, then $e > 0$ and $-e - \epsilon < 0 \implies -e - \epsilon - \bar{\xi}_i < 0$. Hence $\bar{\alpha}_i^* = 0$ by (D2, C2), then $\bar{\alpha}_i + \bar{\beta}_i = C$ by (S3), thus, $0 \leq \bar{\alpha}_i \leq C$. Also $\bar{\alpha}_i + \bar{\beta}_i = C$ implies at least one of $\bar{\alpha}_i > 0$, $\bar{\beta}_i > 0$ is held, thus $\bar{\xi}_i = 0$ by (C1, C3).

3. If $e = -\epsilon$, then $e < 0$ and $e - \epsilon < 0 \implies e - \epsilon - \bar{\xi}_i < 0$. Hence $\bar{\alpha}_i = 0$ by (D1, C1), then $\bar{\alpha}_i^* + \bar{\beta}_i = C$ by (S3), thus, $0 \leq \bar{\alpha}_i^* \leq C$. Also $\bar{\alpha}_i^* + \bar{\beta}_i = C$ implies at least one of $\bar{\alpha}_i^* > 0$, $\bar{\beta}_i > 0$ is held, thus $\bar{\xi}_i = 0$ by (C2, C3)

4. If $e > \epsilon > 0 \implies \begin{cases} e - \epsilon > 0, \\ (-e) - \epsilon < 0 \end{cases}$. $(-e) - \epsilon < 0 \implies (-e) - \epsilon - \bar{\xi}_i < 0$, then $\bar{\alpha}_i^* = 0$ by (D2, C2). Also $e - \epsilon > 0$ and (P1) $\implies \bar{\xi}_i > 0$, then $\bar{\beta}_i = 0$ by (C3). These conditions and (S3) implies $\bar{\alpha}_i = C > 0$ and then $\bar{\xi}_i = e + \epsilon$ by (C1).

5. If $e < -\epsilon \implies \begin{cases} e - \epsilon < 0, \\ (-e) - \epsilon > 0 \end{cases}$. $e - \epsilon < 0 \implies e - \epsilon - \bar{\xi}_i < 0$, then $\bar{\alpha}_i = 0$ by (D1, C1). Also $-e - \epsilon > 0$ and (P2) $\implies \bar{\xi}_i > 0$, then $\bar{\beta}_i = 0$ by (C3). These conditions and (S3) implies $\bar{\alpha}_i^* = C > 0$ and then $\bar{\xi}_i = -(e + \epsilon)$ by (C2).

(d)

1. As in (b) shows, the dual problem maximizes $\sum_{i=1}^m (\alpha_i - \alpha_i^*) y_i - \frac{1}{2} \sum_{0 \leq i,j \leq m} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) x_i^T x_j - \epsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*)$. The kernel function is $k(x_i, x_j) = x_i^T x_j$ for training data x_i, x_j.

2. Since strong duality holds in this problem, the primal optimal solution $(\bar{w}, \bar{b}, \bar{\xi})$ and dual optimal solution $(\bar{\alpha}, \bar{\alpha}^*, \bar{\beta})$ satisfies that $\bar{w} = \sum_{i=1}^m (\bar{\alpha}_i - \bar{\alpha}_i^*) x_i$. For a new data x, we can then reformulate hypothesis $f(x) = w^T x + b$ as:

$$f(x) = \sum_{i=1}^m (\bar{\alpha}_i - \bar{\alpha}_i^*)(x_i^T x) + \bar{b}$$

Problem 4

(a)

$(\implies)$ If $(\bar{\mathbf{w}}, \bar{b}, \bar{\xi})$ optimal, we construct $(\bar{\mathbf{u}}, \bar{\mathbf{v}})$ as such:

$$(\bar{u}_j, \bar{v}_j) = \begin{cases} (\bar{w}_j, 0), & \bar{w}_j \geq 0 \\ (0, -\bar{w}_j), & \bar{w}_j < 0, \end{cases} \quad \forall i = 1, ..., m.$$

Hence $\bar{\mathbf{w}} = \bar{\mathbf{u}} - \bar{\mathbf{v}}$, and note that $(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \bar{b}, \bar{\xi})$ are feasible solution in minimizing f. Since $\sum_{j=1}^{m}(\bar{u}_j + \bar{v}_j) = \sum_{j=1}^{m} |\bar{w}_j| = \|\bar{\mathbf{w}}\|_1$, we have:

$$f(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \bar{b}, \bar{\xi}) = \sum_{i=1}^{m}(\bar{u}_j + \bar{v}_j) + \sum_{i=1}^{N} C_i \bar{\xi}_i = \|\bar{\mathbf{w}}\|_1 + \sum_{i=1}^{N} C_i \bar{\xi}_i.$$

For all other feasible $(\tilde{\mathbf{u}}, \tilde{\mathbf{v}}, \tilde{b}, \tilde{\xi})$, now let $\tilde{\mathbf{w}} = \tilde{\mathbf{u}} - \tilde{\mathbf{v}}$. Note they are feasible in minimizing $\|\mathbf{w}\|_1 + \sum_{i=1}^{N} C_i \xi_i$, and:

$$\|\tilde{\mathbf{w}}\|_1 = \sum_{j=1}^{m} |\tilde{u}_j - \tilde{v}_j| \leq \sum_{j=1}^{m}(\tilde{u}_j + \tilde{v}_j)$$

Hence,

$$f(\tilde{\mathbf{u}}, \tilde{\mathbf{v}}, \tilde{b}, \tilde{\xi}) = \|\tilde{\mathbf{w}}\|_1 + \sum_{i=1}^{N} C_i \tilde{\xi}_i \geq \|\bar{\mathbf{w}}\|_1 + \sum_{i=1}^{N} C_i \bar{\xi}_i = f(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \bar{b}, \bar{\xi}).$$

This shows $(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \bar{b}, \bar{\xi})$ is optimal. $\square$

($\Longleftarrow$) Let $(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \bar{b}, \bar{\xi})$ be optimal, with $\bar{\mathbf{w}} = \bar{\mathbf{u}} - \bar{\mathbf{v}}$. First, note that $(\bar{\mathbf{w}}, \bar{b}, \bar{\xi})$ is feasible in minimizing $\|\mathbf{w}\|_1 + \sum_{i=1}^{N} C_i \xi_i$, and

$$\begin{aligned}
\|\bar{\mathbf{w}}\|_1 + \sum_{i=1}^{N} C_i \bar{\xi}_i &= \sum_{j=1}^{m} |\bar{u}_j - \bar{v}_j| + \sum_{i=1}^{N} C_i \bar{\xi}_i \\
&\leq \sum_{j=1}^{m} |\bar{u}_j| + |-\bar{v}_j| + \sum_{i=1}^{N} C_i \bar{\xi}_i \\
&= \sum_{j=1}^{m}(\bar{u}_j + \bar{v}_j) + \sum_{i=1}^{N} C_i \bar{\xi}_i
\end{aligned}$$

For all feasible $(\tilde{\mathbf{w}}, \tilde{b}, \tilde{\xi})$, let $(\tilde{u}_j, \tilde{v}_j) = \begin{cases} (\tilde{w}_j, 0), & \text{if } \tilde{w}_j \geq 0, \\ (0, -\tilde{w}_j), & \text{if } \tilde{w}_j < 0. \end{cases}$. We have,

$$\begin{aligned}
\|\tilde{\mathbf{w}}\|_1 + \sum_{i=1}^{N} C_i \tilde{\xi}_i &= \sum_{j=1}^{m} |\tilde{u}_j - \tilde{v}_j| + \sum_{i=1}^{N} C_i \tilde{\xi}_i \\
&= \sum_{j=1}^{m}(\tilde{u}_j + \tilde{v}_j) + \sum_{i=1}^{N} C_i \tilde{\xi}_i \\
&\geq \sum_{j}(\bar{u}_j + \bar{v}_j) + \sum_{i=1}^{N} C_i \bar{\xi}_i \\
&\geq \|\bar{\mathbf{w}}\|_1 + \sum_{i=1}^{N} C_i \bar{\xi}_i
\end{aligned}$$

Hence, $(\bar{\mathbf{w}}, \bar{b}, \bar{\xi})$ is optimal. $\square$

(b)

The Lagrangian is :

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

(c)

It suffices to show $f, g_i^1, g_i^2, g_j^3, g_j^4$ are convex function to show Slater's condition are satisfied.

Suppose $(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi})$ and $(\mathbf{u}', \mathbf{v}', b', \boldsymbol{\xi}')$ are distinct primal feasible solution and $t \in [0, 1]$. One can observed that these functions all satisfy that

$$
\begin{aligned}
h(t\mathbf{u} + (1-t)\mathbf{u}', t\mathbf{v} + (1-t)\mathbf{v}', tb + (1-t)b', t\boldsymbol{\xi} + (1-t)\boldsymbol{\xi}') &= th(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}) + (1-t)h(\mathbf{u}', \mathbf{v}', b', \boldsymbol{\xi}') \\
&\le th(\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}) + (1-t)h(\mathbf{u}', \mathbf{v}', b', \boldsymbol{\xi}'),
\end{aligned}
$$

for $h(\cdot) \in \{f, g_i^1, g_i^2, g_j^3, g_j^4\}$. Hence, these functions are all convex and Slater's condition holds.

(d)

(i)

Take partial derivatives over variables of primal problem to Lagrangian:

$$\nabla_{\mathbf{u}} L = \mathbf{1}^T - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i^T - \boldsymbol{\mu}^T \tag{4}$$

$$\nabla_{\mathbf{v}} L = \mathbf{1}^T + \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i^T - \boldsymbol{\nu}^T \tag{5}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \alpha_i y_i \tag{6}$$

$$\frac{\partial L}{\partial \xi_i} = C_i - \alpha_i - \beta_i, i = 1, ..., N \tag{7}$$

If one of (4), (5), (6) not equals to zero, one can follow the gradient to decrease L, thus $\theta(\alpha, \beta, \boldsymbol{\mu}, \boldsymbol{\nu})- \infty$ in these cases. Otherwise, $C_i = \alpha_i + \beta_i \quad \forall i = 1, \ldots, N$ should hold to minimize L. Hence, these equations hold:

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

$$\alpha_i + \beta_i = C_i, \forall i = 1, ..., N. \tag{9}$$

In this case:

$$
\begin{aligned}
&\theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu}) \\
&\quad= \mathbf{1}^T(\mathbf{u}+\mathbf{v}) + \sum_{i=1}^{N} \alpha_i \left(1 - y_i((\mathbf{u}-\mathbf{v})^T\mathbf{x}_i + b)\right) - (\mathbf{1}^T\mathbf{u} - \sum_{i=1}\alpha_i y_i \mathbf{x}_i^T\mathbf{u}) - (\mathbf{1}^T\mathbf{v} + \sum_{i=1}\alpha_i y_i \mathbf{x}_i^T\mathbf{v}) \\
&\quad= \sum_{i=1}^{N} \alpha_i - b\sum_{i=1}^{N}\alpha_i y_i \\
&\quad= \sum_{i=1}^{N} \alpha_i.
\end{aligned}
$$

(ii)

($\Longrightarrow$) If stationary condition holds $(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = \arg\min_{(\mathbf{u}',\mathbf{v}',b',\boldsymbol{\xi}')} L(\mathbf{u}',\mathbf{v}',b',\boldsymbol{\xi}',\alpha,\beta,\mu,\nu)$, their corresponding first derivatives should equal to zero, (8), (9) are satisfied.

($\Longleftarrow$) If (8), (9) are satisfied, $(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})$ set the first derivatives to zero. Since L is convex, this implies $(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi})$ is a minimizer, $(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = \arg\min_{(\mathbf{u}',\mathbf{v}',b',\boldsymbol{\xi}')} L(\mathbf{u}',\mathbf{v}',b',\boldsymbol{\xi}',\alpha,\beta,\mu,\nu)$, stationary condition holds.

(e)

From (d), we have seen that the dual optimal solution should satisfy (8) and (9), and $\theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu}) = \sum_{i=1}^{N}\alpha_i$.

From (8), and dual feasibility constraint, $\boldsymbol{\mu}\geq 0, \boldsymbol{\nu}\geq 0$, we have:

$$
-\mathbf{1} \leq \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \leq \mathbf{1}.
$$

From (9), and dual feasibility constraint $\alpha_i \geq 0, \beta_i \geq 0, i = 1,...,N$, we only need to consider

$$
0 \leq \alpha_i \leq C_i \quad i = 1,...,N.
$$

Combining these results, we have the simplified dual problem as described.

(f)

Suppose $(\bar{\mathbf{u}},\bar{\mathbf{v}},\bar{b},\bar{\boldsymbol{\xi}})$ and $(\bar{\alpha},\bar{\beta},\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\nu}})$ are primal and dual optimal respectively. Since Slater's condition holds, the problem has zero gap, optimality implies these KKT conditions hold:

Stationary

$$
\begin{aligned}
&\text{(S1)} \quad \sum_{i=1}^{N} \bar{\alpha}_i y_i = 0, \\
&\text{(S2)} \quad \bar{\boldsymbol{\mu}} = \mathbf{1} - \sum_{i=1}^{N} \bar{\alpha}_i y_i \mathbf{x}_i, \\
&\text{(S3)} \quad \bar{\boldsymbol{\nu}} = \mathbf{1} + \sum_{i=1}^{N} \bar{\alpha}_i y_i \mathbf{x}_i, \\
&\text{(S4)} \quad \bar{\alpha}_i + \bar{\beta}_i = C_i, && \forall i = 1, \ldots, m
\end{aligned}
$$

Complementary Slackness

$$
\begin{aligned}
&\text{(C1)} \quad \bar{\alpha}_i \left(1 - \bar{\xi}_i - y_i((\bar{\mathbf{u}} - \bar{\mathbf{v}})^T \mathbf{x}_i + \bar{b})\right) = 0, && \forall i = 1, \ldots, N \\
&\text{(C2)} \quad \bar{\beta}_i \bar{\xi}_i = 0, && \forall i = 1, \ldots, N \\
&\text{(C3)} \quad \bar{\boldsymbol{\mu}}^T \bar{\mathbf{u}} = 0 \\
&\text{(C4)} \quad \bar{\boldsymbol{\nu}}^T \bar{\mathbf{v}} = 0
\end{aligned}
$$

Feasibility

$$
\begin{aligned}
&\text{(P1)} \quad 1 - \bar{\xi}_i - y_i((\bar{\mathbf{u}} - \bar{\mathbf{v}})^T \mathbf{x}_i + \bar{b}) \leq 0, && \forall i = 1, \ldots, N \\
&\text{(P2)} \quad -\bar{\xi}_i \leq 0, && \forall i = 1, \ldots, N \\
&\text{(P3)} \quad -\bar{u}_j \leq 0, && \forall j = 1, \ldots, m \\
&\text{(P4)} \quad -\bar{v}_j \leq 0, && \forall j = 1, \ldots, m
\end{aligned}
$$

$$
\begin{aligned}
&\text{(D1)} \quad \bar{\alpha}_i \geq 0, && \forall i = 1, \ldots, N \\
&\text{(D2)} \quad \bar{\beta}_i \geq 0, && \forall i = 1, \ldots, N \\
&\text{(D3)} \quad \bar{\mu}_j \geq 0, && \forall j = 1, \ldots, m \\
&\text{(D3)} \quad \bar{\nu}_j \geq 0, && \forall j = 1, \ldots, m.
\end{aligned}
$$

(g)

Rewrite the target function as

$$
\underbrace{\|\mathbf{w}\|_1}_{L^1 \text{ regularization}} + \sum_{i=1}^{N} C_i \underbrace{\max\{1 - (y_i \mathbf{w}^T \mathbf{x}_i + b), 0\}}_{\text{hinge loss}},
$$

Let $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2$, where $\mathbf{w}_1 \in span\{\mathbf{x}_i\}, \mathbf{w}_2$ is orthogonal to $span\{\mathbf{x}_i\}$.

For the hinge loss part, since $\mathbf{w}_2^T \mathbf{x}_i = 0$, the $\mathbf{w}_2$ part did not contribute to this term.

And by triangle inequality, the upper bound of the L^1 regularization term is:

$$\|\mathbf{w}\|_1 \leq \|\mathbf{w}_1\|_1 + \|\mathbf{w}_2\|_1.$$

Since $\|\mathbf{w}_2\|_1 \geq 0$, the orthogonal part only makes positive contribution to the regularization term.

To minimize the target function, it's equivalent to minimize its upper bound. Hence $\mathbf{w}_2$ should be set to zero, which implies that optimal $\bar{\mathbf{w}}$ is a linear combination of $\mathbf{x}_i$.

Problem 5

1. The Lagrangian function is :

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

2. It suffices to show that Slater's conditions hold. In this case, to check if $f, g_{1,i} g_{2,i}, g_3, \forall i = 1, ..., N$ are all convex. Since $f, g_{2,i}, g_3$ are all linear, they are convex.

   For $g_{1,i}$, suppose $(\rho, \boldsymbol{\mu}, \boldsymbol{\xi})$ and $(\rho', \boldsymbol{\mu}', \boldsymbol{\xi}')$ are all primal feasible, and $t \in [0, 1]$, then:

$$\begin{aligned} g_{1,i}(t\rho + &(1-t)\rho', t\boldsymbol{\mu} + (1-t)\boldsymbol{\mu}', t\boldsymbol{\xi} + (1-t)\boldsymbol{\xi}') \\ &= \|\mathbf{x}_i - (t\boldsymbol{\mu} + (1-t)\boldsymbol{\mu}')\|^2 - t\rho - (1-t)\rho' - t\boldsymbol{\xi}_i - (1-t)\boldsymbol{\xi}'_i \\ &\leq t\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 + (1-t)\|\mathbf{x}_i - \boldsymbol{\mu}'\|^2 - t\rho - (1-t)\rho' - t\boldsymbol{\xi}_i - (1-t)\boldsymbol{\xi}'_i \quad \text{(triangle inequality)} \\ &= g_{1,i}\left(t\rho, t\boldsymbol{\mu}, t\boldsymbol{\xi}\right) + g_{1,i}\left((1-t)\rho', (1-t)\boldsymbol{\mu}', (1-t)\boldsymbol{\xi}'\right), \end{aligned}$$

   which is convex.

3.

$$\frac{\partial L}{\partial \rho} = 1 - \sum_{i=1}^{N}\alpha_i - \gamma, \tag{10}$$

$$\nabla_\mu L = \sum_{i=1}^{N}\alpha_i \cdot 2(\mu - \mathbf{x}_i), \tag{11}$$

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{\nu}C_i - \alpha_i - \beta_i \quad \forall i = 1, \ldots, N. \tag{12}$$

   If (10), (11) be zero, $\frac{C_i}{\nu} = \alpha_i + \beta_i \forall i = 1, ..., N$ to minimize L, otherwise $L = -\infty$. Also (11)$= 0$ implies that $\mu = \frac{\sum_{i=1}^{N}\alpha_i\mathbf{x}_i}{\sum_{i=1}^{N}\alpha_i}$, hence,

$$\theta(\alpha, \beta, \gamma) = \sum_{i=1}^{N}\alpha_i\|\mathbf{x}_i - \mu\|^2 = \sum_{i=1}^{N}\alpha_i\left\|\mathbf{x}_i - \frac{\sum_{i=1}^{N}\alpha_i\mathbf{x}_i}{\sum_{i=1}^{N}\alpha_i}\right\|^2.$$

4. by 3., $\theta(\alpha,\beta,\gamma)$ can be further simplified as:

$$\begin{aligned}
\theta(\alpha,\beta,\gamma) &= \sum_{i=1}^{N} \alpha_i \|\mathbf{x}_i - \mu\|^2 \\
&= \sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x}_i - \mu^T \mathbf{x}_i + \mu^T(\mu - \mathbf{x}_i) \\
&= \|\alpha\|_1 \left(\sum_{i=1}^{N} \hat{\alpha}_i \|\mathbf{x}_i\|^2 - \sum_{i=1}^{N} \hat{\alpha}_i (\sum_{j=1}^{N} \hat{\alpha}_j \mathbf{x}_j^T) \mathbf{x}_i \right) \\
&= \|\alpha\|_1 \left(\sum_{i=1}^{N} \hat{\alpha}_i \|\mathbf{x}_i\|^2 - \sum_{0 \le i,j \le N} \hat{\alpha}_i \hat{\alpha}_j \mathbf{x}_i^T \mathbf{x}_j \right).
\end{aligned}$$

(10) $= 0$ and $\gamma \geq 0$ implies that:

$$\sum_{i=1}^{N} \alpha_i = 1 - \gamma \leq 1.$$

Lastly, (12)$= 0$ and $\alpha_i \geq 0, \beta_i \geq 0, \quad \forall i = 1, ..., N$ implies that:

$$0 \leq \alpha_i \leq \frac{C_i}{\nu}, i = 1, ..., N.$$

Thus the dual problem can be simplified as described.

5. Since this case has zero gap between primal problem and dual solution, KKT conditions hold. They are:

Stationary

$$\begin{aligned}
&\text{(S1)} \quad \sum_{i=1}^{N} \bar{\alpha}_i = 1 - \bar{\gamma}, \\
&\text{(S2)} \quad \sum_{i=1}^{N} \bar{\alpha}_i (\mathbf{x}_i - \bar{\boldsymbol{\mu}}) = 0, \\
&\text{(S3)} \quad \bar{\alpha}_i + \bar{\beta}_i = \frac{C_i}{\nu}, \qquad \forall i = 1, \ldots, m
\end{aligned}$$

Complementary Slackness

$$\begin{aligned}
&\text{(C1)} \quad \bar{\alpha}_i \left(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \bar{\rho} - \bar{\xi}_i \right) = 0, \qquad \forall i = 1, \ldots, N \\
&\text{(C2)} \quad \bar{\beta}_i \bar{\xi}_i = 0, \qquad \forall i = 1, \ldots, N \\
&\text{(C3)} \quad \bar{\gamma} \bar{\rho} = 0
\end{aligned}$$

Feasibility

$$
\begin{aligned}
&\text{(P1)} \quad \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \bar{\rho} - \bar{\xi}_i \le 0, && \forall i = 1, \dots, N \\
&\text{(P2)} \quad -\bar{\xi}_i \le 0, && \forall i = 1, \dots, N \\
&\text{(P3)} \quad -\bar{\rho} \le 0, &&
\end{aligned}
$$

$$
\begin{aligned}
&\text{(D1)} \quad \bar{\alpha}_i \ge 0, && \forall i = 1, \dots, N \\
&\text{(D2)} \quad \bar{\beta}_i \ge 0, && \forall i = 1, \dots, N \\
&\text{(D3)} \quad \bar{\gamma} \ge 0. &&
\end{aligned}
$$

(a) By (S2), $\sum_{i=1}^N \bar{\alpha}_i \mathbf{x}_i = \sum_{i=1}^N \bar{\alpha}_i \bar{\boldsymbol{\mu}} = \|\bar{\alpha}\|_1 \bar{\boldsymbol{\mu}}$.

(b) By (P1), (P2) We have

$$\bar{\xi}_i \ge \max\left\{\|\mathbf{x}_i - \bar{\mu}\|^2 - \bar{\rho}, 0\right\}.$$

Plug-in the target function in primal problem, since $\bar{\rho}$ is the minimizer, it can be formulated as:

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

(c) Rewrite the result of (b), we have

$$\bar{\rho} = \arg\min_{\rho \ge 0} \left\{\rho + \frac{1}{\nu} \sum_{i:\|\mathbf{x}_i - \bar{\mu}\|^2 > \rho} C_i \left(\|\mathbf{x}_i - \bar{\mu}\|^2 - \rho\right)\right\}.$$

Let $\rho_1 = \min\left\{\rho \ge 0 : \sum_{i:\|\mathbf{x}_i - \bar{\mu}\|^2 > \rho} C_i < \nu\right\}$, and $\bar{S} = \{i : \|\mathbf{x}_i - \bar{\mu}\|^2 > \bar{\rho}\}, S_1 = \{i : \|\mathbf{x}_i - \bar{\mu}\|^2 > \rho_1\}, S_D = \{i : \max(\bar{\rho}, \rho_1) > \|\mathbf{x}_i - \bar{\mu}\|^2 \ge \min(\bar{\rho}, \rho_1)\}$.

First we have:

$$\bar{\rho} + \frac{1}{\nu} \sum_{i \in \bar{S}} C_i (\|\mathbf{x}_i - \bar{\mu}\|^2 - \bar{\rho}) \le \rho_0 + \frac{1}{\nu} \sum_{i \in S_1} C_i (\|\mathbf{x}_i - \bar{\mu}\|^2 - \rho_0)$$

Suppose $\bar{\rho} > \rho_1$, note $S_1 = \bar{S} + S_D$ and $\bar{\rho} > \|\mathbf{x}_i - \bar{\mu}\|^2 \ge \rho_1 \quad \forall i \in S_D$, we have:

$$\bar{\rho} + \frac{1}{\nu} \sum_{i \in \bar{S}} C_i (\|\mathbf{x}_i - \bar{\mu}\|^2 - \bar{\rho}) \le \rho_1 + \frac{1}{\nu} \sum_{i \in \bar{S}} C_i \|\mathbf{x}_i - \bar{\mu}\|^2 + \frac{1}{\nu} \sum_{i \in S_D} C_i \|\mathbf{x}_i - \bar{\mu}\|^2 - \frac{\rho_1}{\nu} \sum_{i \in S_1} C_i.$$

$$
\begin{aligned}
\bar{\rho} - \rho_1 &\le \frac{1}{\nu} \sum_{i \in S_D} C_i \|\mathbf{x}_i - \bar{\mu}\|^2 + \frac{\bar{\rho}}{\nu} \sum_{i \in \bar{S}} C_i - \frac{\rho_1}{\nu} \sum_{i \in S_1} C_i \\
&\le \frac{\bar{\rho}}{\nu} \sum_{i \in S_D} C_i + \frac{\bar{\rho}}{\nu} \sum_{i \in \bar{S}} C_i - \frac{\rho_1}{\nu} \sum_{i \in S_1} C_i \\
&= \frac{(\bar{\rho} - \rho_1)}{\nu} \sum_{i \in S_1} C_i
\end{aligned}
$$

Since $\sum_{i\in S_1} C_i < \nu$, $\frac{1}{\nu}\sum_{i\in S_1} C_i < 1$, this leads to contradiction. Hence $\bar{\rho} \le \rho_1$. $\square$
Let $\rho_2 = \min\left\{\rho \ge 0 : \sum_{i:\|\mathbf{x}_i-\bar{\mu}\|^2>\rho} C_i \le \nu\right\}$, and define $S_2 = \{i : \|\mathbf{x}_i - \bar{\mu}\|^2 > \rho_2\}$, $S_c' = \{i : \max(\bar{\rho}, \rho_2) > \|\mathbf{x}_i - \bar{\mu}\|^2 \ge \min(\bar{\rho}, \rho_2)\}$.
Suppose $\bar{\rho} < \rho_2$, note $\bar{S} = S_2 + S_c'$ and $\rho_2 > \|\mathbf{x}_i - \bar{\mu}\|^2 \ge \bar{\rho} \quad \forall i \in S_c'$ we have:

$$\bar{\rho} + \frac{1}{\nu}\sum_{i\in S_2} C_i\|\mathbf{x}_i - \bar{\mu}\|^2 + \frac{1}{\nu}\sum_{i\in S_c'} C_i\|\mathbf{x}_i - \bar{\mu}\|^2 - \frac{\bar{\rho}}{\nu}\sum_{i\in\bar{S}} C_i \le \rho_2 + \frac{1}{\nu}\sum_{i\in S_2} C_i(\|\mathbf{x}_i - \bar{\mu}\|^2 - \rho_2).$$

$$\begin{aligned}
\rho_2 - \bar{\rho} &\ge \frac{1}{\nu}\sum_{i\in S_D'} C_i\|\mathbf{x}_i - \bar{\mu}\|^2 - \frac{\bar{\rho}}{\nu}\sum_{i\in\bar{S}} C_i + \frac{\rho_2}{\nu}\sum_{i\in S_2} C_i \\
&\ge \frac{\bar{\rho_2}}{\nu}\sum_{i\in S_D} C_i - \frac{\bar{\rho}}{\nu}\sum_{i\in\bar{S}} C_i + \frac{\rho_2}{\nu}\sum_{i\in S_2} C_i \\
&= \frac{\rho_2 - \bar{\rho}}{\nu}\sum_{i\in\bar{S}} C_i
\end{aligned}$$

Since $\bar{\rho} < \rho_2$, $\frac{1}{\nu}\sum_{i\in\bar{S}} C_i > 1$, this also leads to contradiction. Hence $\bar{\rho} \ge \rho_2$.

Combining these results, we have:

$$\min\left\{\rho \ge 0 : \sum_{i:\|\mathbf{x}_i-\bar{\mu}\|^2>\rho} C_i \le \nu\right\} \le \bar{\rho} \le \min\left\{\rho \ge 0 : \sum_{i:\|\mathbf{x}_i-\bar{\mu}\|^2>\rho} C_i < \nu\right\} \square$$

(d) If $\bar{\xi}_i > 0$, then $\bar{\beta}_i = 0$ by (C2) and $\bar{\alpha}_i = \frac{C_i}{\nu} > 0$ by (S3). This and (C1) implies that $\|\mathbf{x}_i - \bar{\mu}\|^2 - \bar{\rho} - \bar{\xi}_i = 0$, that is $\bar{\xi}_i = \|\mathbf{x}_i - \bar{\mu}\|^2 - \bar{\rho}, \quad \forall i = 1, ..., N$. Thus $\bar{\xi}_i = \max\{\|\mathbf{x}_i - \bar{\mu}\|^2 - \bar{\rho}, 0\}$.

(e) If $\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \bar{\rho}$, by (d) $\bar{\xi}_i > 0$, then $\bar{\beta}_i = 0$ by (C2), finally $\bar{\alpha}_i = C_i/\nu$ by (S3).
If $\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 < \bar{\rho}$, by (d) $\bar{\xi}_i = 0$. Thus $\bar{\alpha}_i = 0$ by (C1).
If $\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 = \bar{\rho}$, by (d) $\bar{\xi}_i = 0$. Thus $0 \le \bar{\alpha}_i \le C_i/\nu$ by (S3) and (D2).

6. Suppose $C_i = 1/n$ for $i = 1, ..., N$. The objective function becomes:

$$\rho + \frac{1}{n\nu}\sum_{i=1}^{N} \xi_i.$$

By 5., the optimal $\bar{\xi}_i = \max\{\|\mathbf{x}_i - \bar{\mu}\|^2 - \bar{\rho}, 0\}$, only takes positive values when $\|\mathbf{x}_i - \bar{\mu}\|^2 - \bar{\rho} > 0$. In this case, $1/\nu$ is the penalty to the target function from those data points outside of the hypersphere.