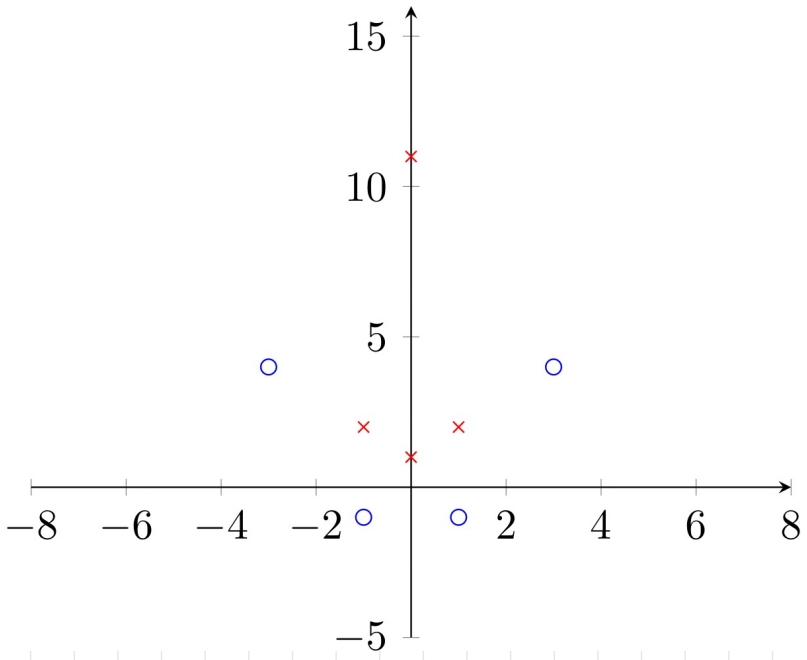
Jonathan Nusantara jan 26 5

Problem 1. (15 points). SVM's obtain *non-linear* decision boundaries by mapping the feature vectors $\bar{X} \in \mathbb{R}^d$ to a possibly high dimensional space via a function $\phi : \mathbb{R}^d \to \mathbb{R}^m$, and then finding a linear decision boundary in the new space.

We also saw that to implement SVM, it suffices to know the kernel function $K(\bar{X}_i, \bar{X}_j) = \phi(\bar{X}_i) \cdot \phi(\bar{X}_j)$, without even explicitly specifying the function ϕ.

Recall **Mercer's theorem**. K is a kernel function if and only if for any n vectors, $\bar{X}_1, \ldots, \bar{X}_n \in \mathbb{R}^d$, and **any** real numbers $c_1, \ldots, c_n$, $\sum_{i=1}^n \sum_{j=1}^n c_i c_j K(\bar{X}_i, \bar{X}_j) \geq 0$.

1. Prove the following half of Mercer's theorem (which we showed in class). If K is a kernel then $\sum_{i=1}^n \sum_{j=1}^n c_i c_j K(\bar{X}_i, \bar{X}_j) \geq 0$.
2. Let $d = 1$, and $x, y \in \mathbb{R}$. Is the function $K(x, y) = x + y$ a kernel?
3. Let $d = 1$, and $x, y \in \mathbb{R}$. Is $K(x, y) = xy + 1$ a kernel?
4. Suppose $d = 2$, namely the original features are of the form $\bar{X}_i = [\bar{X}^1, \bar{X}^2]$. Show that $K(\bar{X}, \bar{Y}) = (1 + \bar{X} \cdot \bar{Y})^2$ is a kernel function. This is called as **quadratic kernel**.
   (**Hint**: Find a $\phi : \mathbb{R}^2 \to \mathbb{R}^m$ (for some m) such that $\phi(\bar{X}) \cdot \phi(\bar{Y}) = (1 + \bar{X} \cdot \bar{Y})^2$).
5. Consider the training examples $\langle [0, 1], 1\rangle$, $\langle [1, 2], 1\rangle$, $\langle [-1, 2], 1\rangle$, $\langle [0, 11], 1\rangle$, $\langle [3, 4], -1\rangle$, $\langle [-3, 4], -1\rangle$, $\langle [1 - 1], -1\rangle$, $\langle [-1, -1], -1\rangle$. We have plotted the data points below.
   - Is the data **linearly classifiable** in the original 2-d space? If yes, please come up with *any* linear decision boundary that separates the data. If no, please explain why.
   - Is the data linearly classifiable in the feature space corresponding to the quadratic kernel. If yes, please come up with *any* linear decision boundary that separates the data. If no, please explain why.



(1)

$$K(\bar{x}_i, \bar{x}_j) = \bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x}_j)$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \cdot c_j \cdot K(\bar{x}_i, \bar{x}_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \cdot \bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x}_j)$$

$$= \left(\sum_{i=1}^{n} c_i \bar{\phi}(x_i)\right)\left(\sum_{j=1}^{n} c_j \bar{\phi}(\bar{x}_j)\right)$$

$$= \left\| \sum_{i=1}^{n} c_i \bar{\phi}(\bar{x}_i) \right\|_2^2 \geq 0$$

So, $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \cdot c_j \cdot K(\bar{x}_i, \bar{x}_j) \geq 0$

(2) K is a kernel if $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i \cdot c_j \cdot K(\bar{x}_i, \bar{x}_j) \geq 0$

is $K(x,y) = x+y$ a kernel?

Let $n=1$, $x = x_i$ and $y = x_j$,

then by Mercer's theorem in (1), K is kernel if:

$$\sum_{i=1}^{1}\sum_{j=1}^{1} c_i \cdot c_j \cdot K(x_i, x_j) = c_1 \cdot c_1 \cdot \overbrace{K(x_1, x_1)}^{x_1 + x_1} \geq 0$$

If $c_1 = -1$ and $x_1 = -1$:

$(-1)(-1)(-1-1) = -2 < 0$ -> does not satisfy Mercer's Theorem

This proof that $K(x,y) = x+y$ is not a kernel

(3) Is $K(x,y) = xy+1$ a kernel?

If K is kernel, then it can be decomposed into dot product of 2 vectors:

$$K(x,y) = \phi(\bar{x}) \cdot \phi(\bar{y})$$
$$= xy + 1$$
$$= \begin{bmatrix} x & 1 \end{bmatrix} \cdot \begin{bmatrix} y \\ 1 \end{bmatrix}$$
$$= \phi(\bar{x})^T \cdot \phi(\bar{y})$$

$$\phi(\bar{x}) = \begin{bmatrix} x \\ 1 \end{bmatrix} \qquad \phi(\bar{y}) = \begin{bmatrix} y \\ 1 \end{bmatrix}$$

Since we have found $\phi(\bar{x})$ and $\phi(\bar{y})$, it satisfies Mercer's theorem proof in (1) to further proof that $K(x,y) = xy+1$ is a kernel.

4. Suppose $d = 2$, namely the original features are of the form $\bar{X}_i = [\bar{X}^1, \bar{X}^2]$. Show that $K(\bar{X}, \bar{Y}) = (1 + \bar{X} \cdot \bar{Y})^2$ is a kernel function. This is called as **quadratic kernel**.
(**Hint**: Find a $\phi : \mathbb{R}^2 \to \mathbb{R}^m$ (for some m) such that $\phi(\bar{X}) \cdot \phi(\bar{Y}) = (1 + \bar{X} \cdot \bar{Y})^2$).

$$K(\bar{X}, \bar{Y}) = (1 + \bar{X} \cdot \bar{Y})^2 \quad \rightarrow \bar{X}_i = [X_1, X_2] \ \& \ \bar{Y}_i = [Y_1, Y_2]$$

$$= (1 + X_1 Y_1 + X_2 Y_2)^2$$

$$= (1 + X_1 Y_1 + X_2 Y_2)(1 + X_1 Y_1 + X_2 Y_2)$$

$$= 1 + X_1 Y_1 + X_2 Y_2 + X_1 Y_1 + X_1^2 Y_1^2 + X_1 X_2 Y_1 Y_2 + X_2 Y_2 + X_1 X_2 Y_1 Y_2 + X_2^2 Y_2^2$$

$$= 1 + 2X_1 Y_1 + 2X_2 Y_2 + 2X_1 X_2 Y_1 Y_2 + X_1^2 Y_1^2 + X_2^2 Y_2^2$$

$$= \phi(\bar{X})^T \cdot \phi(\bar{Y}) \quad \leftarrow 1 \times n \cdot n \times 1 \Rightarrow 1 \times 1$$

$$\phi(\bar{X}) = \begin{bmatrix} \sqrt{1} \\ \sqrt{2} X_1 \\ \sqrt{2} X_2 \\ \sqrt{2} X_1 X_2 \\ X_1^2 \\ X_2^2 \end{bmatrix} \qquad \phi(\bar{Y}) = \begin{bmatrix} \sqrt{1} \\ \sqrt{2} Y_1 \\ \sqrt{2} Y_2 \\ \sqrt{2} Y_1 Y_2 \\ Y_1^2 \\ Y_2^2 \end{bmatrix}$$

$$= (\sqrt{1} \times \sqrt{1}) + (\sqrt{2} X_1 \times \sqrt{2} Y_1) + (\sqrt{2} X_2 \times \sqrt{2} Y_2) + (\sqrt{2} X_1 X_2 \times \sqrt{2} Y_1 Y_2) + (X_1^2 Y_1^2) + (X_2^2 Y_2^2)$$

$$= 1 + 2X_1 Y_1 + 2X_2 Y_2 + 2X_1 X_2 Y_1 Y_2 + X_1^2 Y_1^2 + X_2^2 Y_2^2$$

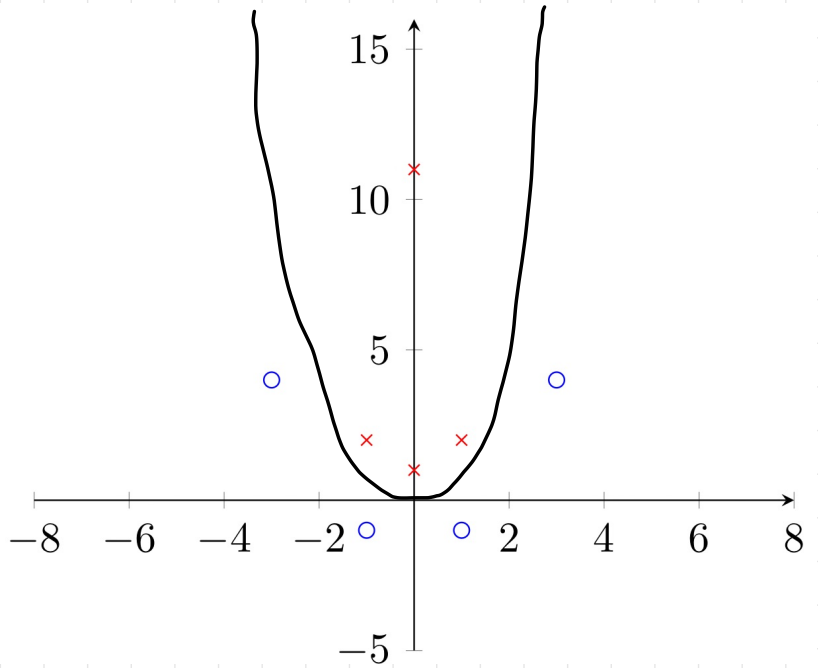
Since we proof to split $K(x, y) = \phi(\bar{X}) \cdot \phi(\bar{Y})$, we can proof Mercer's theorem and that $K(X, Y)$ is a kernel.

⑤ 1. No, it is not linearly classifiable.

There is no single hyperplane w that can correctly classify the data in the original 2-D hyperplane

2. Yes, it is linearly classifiable corresponding to the quadratic kernel.

For example, if the hyperplane looks like this:



This hyperplane has a formula of $y = x^2$ in x-y plane

2 region: $X_2 > (X_1)^2$ or $X_2 < (X_1)^2$

$(X_1)^2 - X_2 < 0$ or $(X_1)^2 - X_2 > 0$

label = 1 or X

label = −1 or O

Problem 2. (10 points). Let $f, h_i, 1 \leq i \leq n$ be real-valued functions and let $\alpha \in \mathbb{R}^n$. Let $L(z, \boldsymbol{\alpha}) = f(z) + \sum_{i=1}^{n} \alpha_i h_i(z)$. In this problem, we will prove that the following two optimization problems are equivalent.

$$\min_z f(z) \quad \text{s.t. } h_i(z) \leq 0,\ i = 1, \ldots, n. \tag{1}$$

$$\min_z \left[\max_{\boldsymbol{\alpha} \geq \mathbf{0}} L(z, \boldsymbol{\alpha}) \right] \tag{2}$$

Let (z^*, α^*) be the solution of (2) and let z_p^* be the solution of (1). Prove that:

$$L(z^*, \alpha^*) = f(z_p^*)$$

Hint: Use the fact that for any z, $\boldsymbol{\alpha} \geq \mathbf{0}$, $L(z^*, \alpha^*) \geq L(z^*, \alpha)$ and $L(z^*, \alpha^*) \leq L(z, \alpha_z)$, where $\alpha_z = \arg\max_{\alpha \geq \mathbf{0}} L(z, \alpha)$.

You may follow the following steps but it is not required as long as your proof is correct.

1. Prove that $L(z^*, \alpha^*) \leq f(z_p^*)$
2. Prove that $L(z^*, \alpha^*) \geq f(z_p^*)$

So if α^* is $\max_{\alpha} L(z, \alpha)$. For any $z, \alpha \geq 0$, $L(z^*, \alpha^*) \geq L(z^*, \alpha)$

Since if we use α^*, it will maximize $L(z^*, \alpha)$.

If z^* is $\min_z \left[\max_{\alpha} L(z, \alpha) \right]$. For any $z, \alpha \geq 0$, $L(z^*, \alpha^*) \leq L(z, \alpha_z)$.

Since if we use z^*, it will minimize $L(z, \alpha^*)$

$$L(z^*, \alpha^*) = f(z^*) + \sum_{i=1}^{n} \alpha_i^* h_i(z^*) \quad \text{s.t. } h_i(z) \leq 0 \ \ i = 1, \ldots, n$$

$$L(z_p^*, \alpha_{z_p^*}) = f(z_p^*) + \sum_{i=1}^{n} \alpha_{z_p^* i} h_i(z_p^*)$$

We know that $L(z^*, \alpha^*) \leq L(z, \alpha_z)$, so

$L(z^*, \alpha^*) \leq L(z_p^*, \alpha_{z_p^*})$ ← because z_p^* may not minimize $L(z, \alpha)$

$$L(z^*, \alpha^*) \leq f(z_p^*) + \sum_{i=1}^{n} \alpha_{z_p^* i} h_i(z_p^*)$$

We also know that $f(z_p^*) + \sum_{i=1}^{n} \alpha_{z_p^* i} h_i(z_p^*) \leq f(z_p^*)$

because by definition $\alpha_{z_p i}^* \geq 0$ and $h_i(z_p^*) \leq 0$ for $i = 1, \ldots, n$

so $\sum_{i=1}^{n} \alpha_{z_p^* i} h_i(z_p^*)$ is negative value that will substract $f(z_p^*)$ term

Thus, we have proven that $L(z^*, \alpha^*) \leq f(z_p^*)$

$$\mathcal{L}(z^*, \alpha^*) = f(z^*) + \sum_{i=1}^{n} \alpha_i^* h_i(z^*)$$

$$\mathcal{L}(z_p^*, \alpha) = f(z_p^*) + \sum_{i=1}^{n} \alpha_i h_i(z_p^*)$$

We know that $\mathcal{L}(z^*, \alpha^*) \geq \mathcal{L}(z^*, \alpha)$, so

$\mathcal{L}(z^*, \alpha^*) \geq \mathcal{L}(z_p^*, \alpha)$ -> because α may not max $\mathcal{L}(z, \alpha)$

$$f(z^*) + \sum_{i=1}^{n} \alpha_i^* h_i(z^*) \geq f(z_p^*) + \sum_{i=1}^{n} \alpha_i h_i(z_p^*)$$

We can also know that z^* satisfies the constraint $h_i(z^*) \leq 0$, $i = 1, \ldots, n$

This is because Lagrange multiplier is based on original + linear combination of constraints.

ex: $f(z^*) + \sum_{i=1}^{n} \alpha_i^* h_i(z^*)$

$f(z^*)$: original function to max or min

$\sum_{i=1}^{n} \alpha_i^* h_i(z^*)$: linear combination of constraint

So, since z^* satisfy the constraint $h_i(z^*) \leq 0$ for $i = 1, \ldots, n$

we also know that $f(z_p^*) \leq f(z^*)$.

This is because z_p^* will minimize (1), as mentioned in the problem.

So, back to the equation we have:

$$f(z^*) + \sum_{i=1}^{n} \alpha_i^* h_i(z^*) \geq f(z_p^*) + \underbrace{\sum_{i=1}^{n} \alpha_i h_i(z_p^*)}_{\leq 0}$$

-> ≤ 0, as $\alpha \geq 0$ and $h_i(z) \leq 0$

So, $f(z^*) + \sum_{i=1}^{n} \alpha_i^* h_i(z^*) \geq f(z_p^*)$, since

$f(z^*) \geq f(z_p^*)$ and $\sum_{i=1}^{n} \alpha_i^* h_i(z^*) \geq \sum_{i=1}^{n} \alpha_i h_i(z_p^*)$ (α will maximize $\mathcal{L}(z, \alpha)$)

=> $\mathcal{L}(z^*, \alpha^*) \geq f(z_p^*)$ is proven

Since we've proven that $\mathcal{L}(z^*, \alpha^*) \leq f(z_p^*)$

and $\mathcal{L}(z^*, \alpha^*) \geq f(z_p^*)$

This means that $\mathcal{L}(z^*, \alpha^*) = f(z_p^*)$

reference : Wikipedia on Lagrange & KKT.

Problem 3. (15 points). In this problem, we derive the dual formulation of the soft-margin SVM problem with $\bar{\xi} = \xi_1, \ldots, \xi_n$.

$$\min_{\bar{w},\bar{\xi}} \frac{1}{2}\|\bar{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \xi_i$$
$$\text{such that} \tag{3}$$
$$1 - \xi_i - y_i(\bar{X}_i \cdot \bar{w} - t) \leq 0,\ i = 1, \ldots, n,$$
$$-\xi_i \leq 0,\ i = 1, \ldots, n.$$

Now we can define $2n$ Lagrangian variables $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_n$, and $\boldsymbol{\beta} = \beta_1, \ldots, \beta_n$ corresponding to these equations and obtain the following Lagrangian.

$$L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\bar{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i(1 - \xi_i - y_i(\bar{X}_i \cdot \bar{w} - t)) - \sum_{i=1}^{n} \beta_i \xi_i \tag{4}$$

The original problem is now equivalent to

$$\min_{\bar{w},t,\bar{\xi}} \left[\max_{\boldsymbol{\alpha} \geq \mathbf{0}, \boldsymbol{\beta} \geq \mathbf{0}} L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \right].$$

For this objective, minmax theorem says that the min max problem is equivalent to the max min problem below. You do not have to prove this, but are encouraged to do so (using the argument given in the discussion earlier). This gives the following problem.

$$\max_{\boldsymbol{\alpha} \geq \mathbf{0}, \boldsymbol{\beta} \geq \mathbf{0}} \left[\min_{\bar{w},t,\bar{\xi}} L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \right].$$

1. For a fixed $\boldsymbol{\alpha}, \boldsymbol{\beta}$ take the gradient of the the Lagrangian with respect to $\bar{w}$, and express $\bar{w}$ in terms of the other variables.
2. Differentiate the Lagrangian with respect to t, and equate to zero to obtain another equation.
3. Differentiate the Lagrangian with respect to ξ_i and show that $\alpha_i + \beta_i = C$ at the optimum. This shows that $\alpha_i \leq C$, since $\beta_i \geq 0$.
4. The expressions from 1, 2, 3 define one of the KKT conditions. Show that under these conditions,
$$\min_{\bar{w},t,\bar{\xi}} L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\bar{X}_i \cdot \bar{X}_j).$$
Basically, use the expressions from 1, 2, and 3 to "cancel out" $\bar{w}, t, \bar{\xi}$ in the Lagrangian.
5. Combine the results above to argue that the following optimization is equivalent to the soft margin SVM we started with.
$$\max_{\boldsymbol{\alpha} \geq \mathbf{0}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\bar{X}_i \cdot \bar{X}_j) \tag{5}$$
$$\text{such that}$$
$$0 \leq \alpha_i \leq C,\ i = 1, \ldots, n.$$

(1) $L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\bar{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i(1 - \xi_i - y_i(\bar{X}_i \cdot \bar{w} - t)) - \sum_{i=1}^{n} \beta_i \xi_i$

$$\nabla_{\bar{w}} L(\bar{w}, t, \bar{\xi}, \alpha, \beta) = \bar{w} + 0 + \frac{d}{d\bar{w}} \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i(\bar{X}_i \cdot \bar{w} - t)) - 0$$
$$= \bar{w} - \sum_{i=1}^{n} \alpha_i \cdot y_i \cdot \bar{X}_i \quad \Rightarrow 0$$

$$\boxed{\bar{w} = \sum_{i=1}^{n} \alpha_i \cdot y_i \cdot \bar{X}_i}$$

② $\nabla_t \mathcal{L}(\bar{w}, t, \bar{\xi}, \alpha, \beta) = 0 + 0 + \sum_{i=1}^{n} \alpha_i y_i - 0 \quad \Rightarrow 0$

$$\boxed{\sum_{i=1}^{n} \alpha_i y_i = 0}$$

③ $\nabla_\xi \mathcal{L}(\bar{w}, t, \bar{\xi}, \alpha, \beta) = 0 + \sum_{i=1}^{n} C - \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \beta_i \quad \Rightarrow 0$

$$\sum_{i=1}^{n} C = \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \beta_i \quad \Rightarrow \quad \boxed{C = \alpha_i + \beta_i}$$

④ $$L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\bar{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i(\bar{X}_i \cdot \bar{w} - t)) - \sum_{i=1}^{n} \beta_i \xi_i$$

$$= \frac{1}{2}\| \sum_{i=1}^{n} \alpha_i \cdot y_i \cdot \bar{x}_i \|_2^2 + (\alpha_i + \beta_i) \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i \xi_i - \sum_{i=1}^{n} \alpha_i y_i (\bar{x}_i \cdot \bar{w} - t) - \sum_{i=1}^{n} \beta_i \xi_i$$

$$= \frac{1}{2}\| \sum_{i=1}^{n} \alpha_i \cdot y_i \cdot \bar{x}_i \|_2^2 + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i (\bar{x}_i \cdot w - t)$$

$$= \frac{1}{2}\left(\sum_{i=1}^{n} \alpha_i \cdot y_i \cdot \bar{x}_i \right)\left(\sum_{j=1}^{n} \alpha_j \cdot y_j \cdot \bar{x}_j \right) + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \left(\bar{x}_i \cdot \left(\sum_{j=1}^{n} \alpha_j y_j \bar{X}_j \right) - t \right)$$

$$= \sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\bar{X}_i \cdot \bar{X}_j) - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\bar{X}_i \cdot \bar{X}_j) + \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{\sum_{i=1}^{n} \alpha_i y_i = 0} t$$

$\mathcal{L}(\bar{w}, t, \bar{\xi}, \alpha, \beta)$ $= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\bar{X}_i \cdot \bar{X}_j)$

Since from ③, $\alpha_i \leq C$ and $\alpha_i \geq 0$, meet KKT constraint criterion. So, optimized and become minimum

$$\min_{\bar{w}, t, \bar{\xi}} L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\bar{X}_i \cdot \bar{X}_j).$$

⑤ Since we know that $\min_{\bar{w}, t, \bar{\xi}} L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\bar{X}_i \cdot \bar{X}_j).$

Then the max-min problem is:

$$\max_{\alpha \geq 0, \beta \geq 0} \min L(\bar{w}, t, \bar{\xi}, \alpha, \beta) \Rightarrow \max_{\alpha \geq 0, \beta \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\bar{x}_i \cdot \bar{x}_j)$$

where since $C = \alpha + \beta$ and $\beta \geq 0$, $\alpha_i \leq C$.

So, the equation above follows $0 \leq \alpha_i \leq C$

Equivalent to the soft margin SVM we started with.