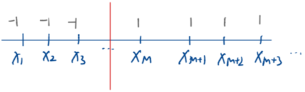
1. (20 points) Consider N "linearly separable" 1D examples $\{(x_n, y_n)\}_{n=1}^{N}$. That is, $x_n \in \mathbb{R}$. Without loss of generality, assume that $x_1 \le x_2 \le \ldots \le x_M < x_{M+1} \le x_{M+2} \le \ldots \le x_N$, $y_n = -1$ for $n = 1, 2, \ldots, M$, and $y_n = +1$ for $n = M+1, M+2, \ldots, N$. Apply hard-margin SVM without transform on this data set. What is the largest margin achieved in terms of $\{(x_n, y_n)\}_{n=1}^{N}$ and M? Prove your answer.

(*Note: This is a "hard-margin" decision stump. Have we mentioned that a decision stump is just a 1D perceptron, and the hard-margin SVM is an extension of the perceptron model? :-)*)



$$\text{margin}(w) = \min_{n=1\sim N} \text{distance}(x_n, w)$$

$$\frac{x_{M+1} - x_M}{2} \quad \#$$

2. (20 points) In some situations, we expect to achieve a smaller margin for the positive examples and a larger margin for the negative examples. For instance, when there are very few negative examples and a lot more positive examples, giving the negative examples a bigger margin could be more robust. Consider an *uneven-margin* SVM that solves

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & (\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \text{ for } y_n = +1 \\ & -(\mathbf{w}^T\mathbf{x}_n + b) \geq \rho \text{ for } y_n = -1. \end{aligned}$$

Derive the (Lagrange) dual problem for the uneven margin SVM above. List your derivation steps.

(*Note: This is a very useful extension of SVM, and more importantly, "forces" you to go through the derivation steps of the class slides to gain a deeper understanding of how to derive a dual problem.*)

$$\mathcal{L}(b, w, \alpha) = \frac{1}{2} w^T w + \sum_{n=1}^{N} a_n (1 - y_n(w^T x_n + b)) + \sum_{n=1}^{N} b_n (\rho + y_n (w^T x_n + b))$$

$$= \frac{1}{2} w^T w + \sum_{n=1}^{N} a_n (1 - w^T x_n - b) + \sum_{n=1}^{N} C_n (\rho - w^T x_n - b)$$

$$\min(\max \mathcal{L}(b, w, \alpha)) \geq \max_{a_n \geq 0} (\min_{b, w} \mathcal{L}(b, w, \alpha))$$

- $\dfrac{\partial \mathcal{L}(b, w, \alpha)}{\partial b} = 0 = -\sum_{n=1}^{N} a_n - \sum_{n=1}^{N} C_n = \sum_{n=1}^{N} (a_n + C_n)$

- $\dfrac{\partial \mathcal{L}(b, w, \alpha)}{\partial w} = 0 = w_i - \sum_{n=1}^{N} \alpha_n x_{n,i} - \sum_{n=1}^{N} c_n x_{n,i}$

$$w_i = \sum_{n=1}^{N} (\alpha_n + c_n) x_{n,i}$$

$$\max_{\substack{a_n \geq 0 \\ \sum a_n + C_n = 0 \\ w = \sum (a_n + C_n) x_n}} \left(\frac{1}{2} w^T w + \sum a_n - a_n w^T x_n - a_n b + \sum C_n \rho - C_n w^T x_n - C_n b \right)$$

↳ if $\mathcal{L}(b, w, \alpha c)$ will be in those constrants if minimized.

$$= \max \left(\frac{1}{2} w^T w + \sum_{n=1}^{N} (a_n + C_n \rho - a_n w^T x_n - C_n w^T x_n) \right)$$

$$= \max \left(\frac{1}{2} w^T w + \sum_{n=1}^{N} (a_n + C_n \rho - w^T w) \right)$$

$$= \max \left(-\frac{1}{2} \sum_{n=1}^{N} \| (a_n + C_n) x_n \|^2 + \sum_{n=1}^{N} (a_n + C_n \rho) \right)$$

$$= \max \left(-\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} (a_n + C_n)(a_m + C_m) x_n x_m + \sum_{n=1}^{N} (\alpha_n + C_n \rho) \right) *$$

2. (20 points) In some situations, we expect to achieve a smaller margin for the positive examples and a larger margin for the negative examples. For instance, when there are very few negative examples and a lot more positive examples, giving the negative examples a bigger margin could be more robust. Consider an *uneven-margin* SVM that solves

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & (\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \text{ for } y_n = +1 \\ & -(\mathbf{w}^T\mathbf{x}_n + b) \geq \rho \text{ for } y_n = -1. \end{aligned}$$

Derive the (Lagrange) dual problem for the uneven margin SVM above. List your derivation steps. (*Note: This is a very useful extension of SVM, and more importantly, "forces" you to go through the derivation steps of the class slides to gain a deeper understanding of how to derive a dual problem.*)

putting two constraints into one $\quad y_n(w^Tx_n+b)\geq 1$, $\quad y_n(w^Tx_n+b)\geq \rho$ $\rightarrow$ $y_n(w^Tx_n+b)+[\![y=-1]\!](1-\rho)\geq 1$

$y_n[(w^Tx_n+b)+[\![y=-1]\!](\rho-1)]\geq 1$

$$\mathcal{L}(b,w,\alpha) = \frac{1}{2}w^Tw + \sum_{n=1}^{N}\alpha_n\left(1-y_n\left(w^Tx_n+b+[\![y=-1]\!](\rho-1)\right)\right)$$

- $\max_{a_n\geq 0} \mathcal{L}(b,w,\alpha) = \infty$, if $y_n(w^Tx_n+b+[\![y=-1]\!](\rho-1))\leq 1$
- $\max_{a_n\geq 0} \mathcal{L}(b,w,\alpha) =$ constant if $y_n(w^Tx_n+b+[\![y=-1]\!](\rho-1)\geq 1$

$$\min_{b,w}\left(\max_{a_n\geq 0}\mathcal{L}(b,w,\alpha)\right) \geq \max_{a_n\geq 0}\left(\min \mathcal{L}(b,w,\alpha)\right)$$

$$\min_{w,b}\left(\max_{\alpha_n>0}\mathcal{L}(w,b,\alpha_n)\right) \geq \max_{\alpha_n>0}\left(\min_{b,w}\left(\frac{1}{2}w^Tw+\sum_{n=1}^{N}\alpha_n\left(1-y_n(w^Tx_n+b)+[\![y=-1]\!](\rho-1)\right)\right)\right)$$

$\min(\mathcal{L}(b,w,\alpha))$ happens when $\frac{\partial\mathcal{L}}{\partial b}=0$ and $\frac{\partial\mathcal{L}}{\partial w}=0$

- $\frac{\partial\mathcal{L}}{\partial b} = 0 = \sum_{n=1}^{N} -y_n a_n$
- $\frac{\partial\mathcal{L}}{\partial w_i} = 0 = w_i - \sum_{n=1}^{N}\alpha_n y_n x_{n,i}$, $\quad w_i = \sum_{n=1}^{N}\alpha_n y_n x_{n,i}$

$$\max_{\substack{a_n\geq 0 \\ \sum y_n a_n = 0 \\ w=\sum a_n y_n x_n}} -\frac{1}{2}\left\|\sum_{n=1}^{N}\alpha_n y_n x_n\right\|^2 + \sum_{n=1}^{N}\left(\alpha_n - y_n b + [\![y=-1]\!](\rho-1)\right)\ \#$$

3. (20 points) Let's call the primal solution to the uneven margin SVM above $(b^*_\rho, \mathbf{w}^*_\rho)$. Assume that Dr. Even has solved the original even-margin SVM

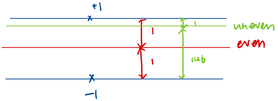
$$\begin{aligned}\min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & (\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \text{ for } y_n = +1 \\ & -(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \text{ for } y_n = -1.\end{aligned}$$

That is, Dr. Even obtained $(b^*_1, \mathbf{w}^*_1)$. Somehow zir boss likes the magic number of 1126 better and wants zir to return $(b^*_{1126}, \mathbf{w}^*_{1126})$ instead. Describe a systematic process to convert $(b^*_1, \mathbf{w}^*_1)$ to $(b^*_{1126}, \mathbf{w}^*_{1126})$ *without* solving the QP problem again. Prove the correctness of your process.

(Note: The result should reinforce your geometric sense of what SVM is doing.)

$$\min_{w,b} \ \frac{1}{2}w^Tw \quad \text{subject to} \quad y_n(w^Tx_n + b) \geq 1$$

$$\min_{w,b} \ \frac{1}{2}w^Tw \quad \text{subject to} \quad \begin{cases} (w^Tx_n + b) \geq 1 & , y = +1 \\ -(w^Tx_n + b) \geq 1126 & , y = -1 \end{cases}$$



even: $(x^+ - x^-) \cdot \frac{w}{\|w\|}$

$= \frac{1}{\|w\|}(x^+w - x^-w)$

$= \frac{1}{\|w\|}((1-b) - (-1-b))$

$= \frac{2}{\|w\|}$

uneven: $(x^+ - x^-) \cdot \frac{W_{1126}}{\|W_{1126}\|}$

$= \frac{1}{\|W_{1126}\|}(W_{1126}x^+ - W_{1126}x^-)$

$= \frac{1}{\|W_{1126}\|}((1-b) - (-1126-b))$

$= \frac{1127}{\|W_{1126}\|}$

Since $\frac{2}{\|W_1\|} = \frac{1127}{\|W_{1126}\|}$, $\|W_{1126}\| = \frac{1127}{2}\|W_1\|$ #

$W_1x^+ + b_1 = W_{1126}x^+ + b_{1126} = 1$

$\Rightarrow b_{1126} = b_1 + (W_1 - W_{1126})x^+$

$= b_1 + -\frac{1125}{2}W_1x^+ = b_1 - \frac{1125}{2}(1-b_1) = \frac{1127}{2}b_1 - \frac{1125}{2}$ #

4. (20 points) Let's call the dual solution to the uneven margin SVM above $\boldsymbol{\alpha}_\rho^*$. Assume that Dr. Even has solved the dual problem of the even-margin SVM and obtained $\boldsymbol{\alpha}_1^*$. Prove or disprove that $\boldsymbol{\alpha}_1^*$ is also an optimal solution of the uneven margin SVM above for any $\rho > 0$.

(Note: Think about how your geometric sense from the previous problem can be extended to algebraic results.)

$\rho > 0$

$$* \min_{w,b}\left(\max_{\alpha_n>0} \mathcal{L}(w,b,\alpha_n)\right) \geq \max_{\alpha_n>0}\left(\min_{b,w}\left(\frac{1}{2}w^Tw + \sum_{n=1}^{N}\alpha_n\left(1-y_n(w^Tx_n+b) + [\![y=-1]\!](\rho-1)\right)\right)\right)$$

$\min(\mathcal{L}_\rho(b,w,\alpha))$ happens when $\frac{\partial \mathcal{L}_\rho}{\partial b}=0$ and $\frac{\partial \mathcal{L}_\rho}{\partial w}=0$

- $\frac{\partial \mathcal{L}_\rho}{\partial b} = 0 = \sum_{n=1}^{N} -y_n a_n$
- $\frac{\partial \mathcal{L}_\rho}{\partial w_i} = 0 = w_{\rho,i} - \sum_{n=1}^{N}\alpha_n y_n x_{n,i}$, $w_{\rho,i} = \sum_{n=1}^{N}\alpha_n y_n x_{n,i}$

$\rho = 1$

$$* \mathcal{L}(b,w,\alpha) = \frac{1}{2}w^Tw + \sum_{n=1}^{N}\alpha_n\left(1-y_n(w^Tx_n+b)\right)$$

$$\frac{\partial \mathcal{L}_1}{\partial b_1} = 0 = \sum_{n=1}^{} -y_n\alpha_n$$

$$\frac{\partial \mathcal{L}_1}{\partial w_{1,i}} = 0 \quad , \quad w_{1,i} = \sum_{n=1}^{N}\alpha_n y_n x_{n,i}$$

→ Both $\frac{\partial \mathcal{L}}{\partial b}$, $\frac{\partial \mathcal{L}}{\partial w_i}$ are dependent of ρ, the optimal solution α_ρ^*

is also the optimal solution for α_1^* #

5. (20 points) For two valid kernels $K_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_1(\mathbf{x})^T \boldsymbol{\phi}_1(\mathbf{x}')$ and $K_2(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_2(\mathbf{x})^T \boldsymbol{\phi}_2(\mathbf{x}')$, consider a kernel function $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') \cdot K_2(\mathbf{x}, \mathbf{x}')$. Prove that K is a valid kernel by deriving a transform function $\boldsymbol{\phi}(\mathbf{x})$ such that

$$K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x}').$$

(Note: The result shows that a multiplication of valid kernels is still a valid kernel. So, for instance, a multi-variate Gaussian kernel is a valid kernel; a polynomial kernel—repeated multiplications of the identity $\boldsymbol{\phi}$—is a valid kernel.)

$$K(x, x') = K_1(x, x') \cdot K_2(x, x')$$

$$= \phi_1(x)^T \phi(x') \cdot \phi_2(x)^T \phi_2(x')$$

$$= \sum_{i=1}^{N} \phi_{1,i}(x) \phi_{1,i}(x') \cdot \sum_{j=1}^{M} \phi_{2,j}(x) \cdot \phi_{2,j}(x')$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{M} (\phi_{1,i}(x) \phi_{2,j}(x)) \cdot (\phi_{1,i}(x') \cdot \phi_{2,j}(x'))$$

$\phi_{3,k} = \phi_{1,i} \cdot \phi_{2,j}$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \phi_{3,k}(x) \cdot \phi_{3,k}(x')$$

$$= \phi_3(x)^T \phi_3(x')$$

$$= K_3(x, x') \#$$

6. (20 points) For any feature transform ϕ from $\mathcal{X}$ to $\mathcal{Z}$, the $\mathcal{Z}$-space distance between two examples $\mathbf{x}$ and $\mathbf{x}'$ is $\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|$, which can be computed with the kernel trick. Consider the degree-2 quadratic kernel $K_2(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^T\mathbf{x}')^2$. For two unit vectors $\mathbf{x}$ and $\mathbf{x}'$, what are their largest and smallest possible distances in the $\mathcal{Z}$ space? Prove your claim.

(Note: This shall sharpen your understanding of the kernel trick.)

$K_2(x, x') = 1 + \sqrt{2}x_1 + \sqrt{2}x_2 \cdots \sqrt{2}x_d + x_1^2 + x_2^2 \cdots x_d^2$

$\|x\| = 1$

$\|\phi(x)\| = \sqrt{\phi^T(x)\phi(x)} = \sqrt{K_2(x)} = \sqrt{(1+x^Tx)^2} = 2$

max $K_2(x, x')$ happens when x, x' in same direction $K_2(x, x') = (1+1)^2 = 4$

min $K_2(x, x')$ happens when x, x' in opposite direction $K_2(x, x') = (1-1)^2 = 0$

$\|\phi(x) - \phi(x')\| = \sqrt{\phi(x)^2 + \phi(x')^2 - 2\phi(x)\phi(x')}$

max $\|\phi(x) - \phi(x')\| = \sqrt{4+4-2\times 0} = 2\sqrt{2}$ #

min $\|\phi(x) - \phi(x')\| = \sqrt{4+4-2\times 4} = 0$ #

7. (20 points) Recall that in the derivation of (one-dimensional) Gaussian kernel, we derived its associated transform

$$\phi(x) = \exp(-x^2) \cdot \left(1, \sqrt{\tfrac{2}{1!}}x, \sqrt{\tfrac{2^2}{2!}}x^2, \dots\right).$$

Let

$$\tilde{\phi}(x) = \left(1, \sqrt{\tfrac{2}{1!}}x, \sqrt{\tfrac{2^2}{2!}}x^2, \dots\right).$$

Prove that $\exp(-x^2) = \frac{1}{\|\tilde{\phi}(x)\|}$.

(Note: The result means $\phi(x)$ can be viewed as a normalized version of $\tilde{\phi}(x)$.)

$$\|\tilde{\phi}(x)\|^2 = \sum_{n=0}^{\infty} \frac{2^n}{n!} x^{2n} = \sum_{n=0}^{\infty} \frac{(\sqrt{2}x)^{2n}}{n!} = \sum_{n=0}^{\infty} \frac{(2x^2)^n}{n!}$$

$$\|\tilde{\phi}(x)\| = \sum_{n=0}^{\infty} \frac{(x^2)^n}{n!}$$

$\downarrow$ Taylor series expansion of $e^{x^2} = \sum_{n=0}^{\infty} \frac{(2x^2)^n}{n!}$

$$\exp(x^2) = \frac{1}{\|\tilde{\phi}(x)\|} \quad \#$$

8. (20 points) Let $\cos(\mathbf{x}, \mathbf{x}')$ measure the cosine of the angle between two non-zero vectors $\mathbf{x}$ and $\mathbf{x}'$ in $\mathbb{R}^d$. The function cos is typically called the cosine similarity between two vectors. Prove or disprove that $\cos(\mathbf{x}, \mathbf{x}')$ is a valid kernel on $\mathcal{X} = \mathbb{R}^d - \{\mathbf{0}\}$.

(*Hint: to prove, you'd better construct its associated transform; to disprove, you may use Mercer's condition to construct a counter-example on positive semi-definiteness.*)

$$K_{ij} = \cos(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|} = \frac{x_i}{\|x_i\|} \cdot \frac{x_j}{\|x_j\|} = \phi_i(x) \cdot \phi_j(x)$$

By considering a vector $C = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \end{bmatrix}$

$$\cos(x, x') = \frac{x^T}{\|x\|} \cdot \frac{x'}{\|x\|} = \frac{x'}{\|x\|} \cdot \frac{x^T}{\|x\|} = \cos(x', x) \rightarrow \text{symmetric}$$

$$C^T K C = \sum_{i=0}^{M} \sum_{j=0}^{N} c_i c_j \cos(x_i \cdot x_j)$$

$$= \sum_{i=0}^{M} \sum_{j=0}^{N} \frac{c_i \cdot c_j}{\|x_i\| \|x_j\|} x_i \cdot x_j$$

$$= \sum_{i=0}^{M} \sum_{j=0}^{N} \left(\frac{c_i}{\|x_i\|} \cdot x_i \cdot \frac{c_j}{\|x_j\|} x_j \right)$$

$$= \left| \sum_{i=0}^{N} \frac{c_i}{\|x_i\|} x_i \right|^2 \geq 0 \rightarrow K \text{ is positive semi-definite}$$

$\rightarrow \cos(x, x')$ is valid kernal.

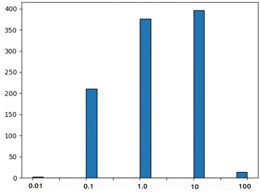
9. (20 points, *) Consider the polynomial kernel $K(\mathbf{x}_n, \mathbf{x}_m) = (1 + \mathbf{x}_n^T \mathbf{x}_m)^Q$, where Q is the degree of the polynomial. Consider a binary classifier of "4" versus "not 4". Among $C \in \{0.1, 1, 10\}$ and $Q \in \{2, 3, 4\}$, count the number of support vectors of the resulting SVM classifier. Which of the (C, Q) combination(s) result in the smallest number of support vectors? Provide your command/code.

```
最小支持向量数量: 629
对应的 (C, Q) 组合: (10, 4)
```

10. (20 points, *) Consider the Gaussian kernel $K(\mathbf{x}_n, \mathbf{x}_m) = \exp\left(-\gamma\|\mathbf{x}_n - \mathbf{x}_m\|^2\right)$. For the binary classification problem of "1" versus "not 1", when fixing $\gamma = 1$, among $C \in \{0.01, 0.1, 1, 10, 100\}$, which value(s) of C result in the lowest E_{out}? Provide your command/code.

```
最小 Eout: 0.5%
对应的 C 值: 1
```

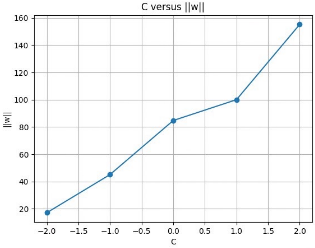
11. (*) Following Problem 10 and consider a validation procedure that randomly samples 200 examples from the training set for validation and leaves the other examples for training g^{-}_{SVM}. Fix $\gamma = 1$ and use the validation procedure to choose the best C among $\{0.01, 0.1, 1, 10, 100\}$ according to E_{val}. If there is a tie of E_{val}, choose the smallest C. Repeat the procedure 1000 times and count the number of times that each C is selected. Plot a bar chart of C versus its selection frequency. Compare the result to that of Problem 10. Describe your findings. Provide your command/code.



We repeat the process 1000 times and calculate every best c for Eval, the most frequent best c would be 10, while 0.01 appears less than 10 times. Since every time Xtrain and Xval are different set, we get different solutions.

In comparison in 5. 10, Xtrain and Xtest is fixed, the best c would never change, best c=1.

12. (20 points, *) Consider the Gaussian kernel $K(\mathbf{x}_n, \mathbf{x}_m) = \exp\left(-\gamma \|\mathbf{x}_n - \mathbf{x}_m\|^2\right)$. For the binary classification problem of "3" versus "not 3", when fixing $\gamma = 1$, vary $C \in \{0.01, 0.1, 1, 10, 100\}$ and calculate $\|\mathbf{w}\|$ of the resulting SVM classifier with the kernel trick. Plot a line chart of C versus $\|\mathbf{w}\|$. Describe your findings. Provide your command/code.



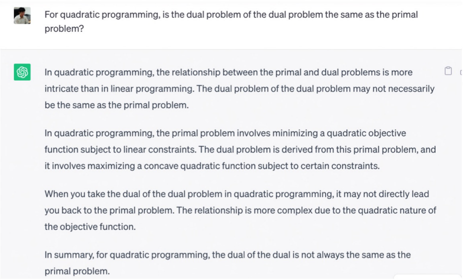
In the plot, when c gets larger, |w| gets larger.
when c gets smaller, |w| gets smaller.

|w| increase, since c indicates the penalty when wrong data occur, |w| indicates the dimension of w, if we want less wrong data, w must be in higher dimension, and might cause overfitting.

When C gets lower, the penalty on wrong data decrease, and algorithm would prefer shorter w.

13. (Bonus 20 points) Derive a simplified Langrange dual problem of the hard-margin SVM dual. Is your Lagrange dual problem of the hard-margin SVM dual the same as the hard-margin SVM primal? Are they "similar" in any sense? Describe your findings.

For your information, this is what chatGPT says.



For quadratic programming, is the dual problem of the dual problem the same as the primal problem?

In quadratic programming, the relationship between the primal and dual problems is more intricate than in linear programming. The dual problem of the dual problem may not necessarily be the same as the primal problem.

In quadratic programming, the primal problem involves minimizing a quadratic objective function subject to linear constraints. The dual problem is derived from this primal problem, and it involves maximizing a concave quadratic function subject to certain constraints.

When you take the dual of the dual problem in quadratic programming, it may not directly lead you back to the primal problem. The relationship is more complex due to the quadratic nature of the objective function.

In summary, for quadratic programming, the dual of the dual is not always the same as the primal problem.

SVM : $\min\limits_{w,b} \left(\max \mathcal{L}(w, b, \alpha_n) \right) \geq \max\limits_{\alpha_n > 0} \left(\min \mathcal{L}(w, b, \alpha_n) \right)$

$$= \max_{\alpha_n > 0} \left(\min_{b,w} \frac{1}{2} w^T w + \sum_{n=1}^{N} \alpha_n \left(1 - y_n (w^T x_n + b) \right) \right)$$

Strong duality ensure that $\min \left(\max \mathcal{L}(b, w, \alpha_n) \right)$

$= \max \left(\min \mathcal{L}(b, w, \alpha_n) \right)$

(convex primal,
feasible primal,
linear constraints)

The dual of the dual problem can be reversed back to the primal.