CS771: Assignment 1

Aniket Sanghi
Roll no: 170110
sanghi@iitk.ac.in

Paramveer Raol
Roll no:170459
paramvir@iitk.ac.in

Sarthak Singhal
Roll no:170635
ssinghal@iitk.ac.in

Mihir Jewalikar
Roll no:170387
mihirsj@iitk.ac.in

Abhishek Bhatia
Roll no:170022
avishek@iitk.ac.in

1 Part 1

The primal problem P2 is given as follows :

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d} \{\tfrac{1}{2}||\mathbf{w}||_2^2 \ + \ C\sum_{i=1}^n \xi_i^2\}$$

$$\text{st. } y^i \cdot \langle \mathbf{w}, \mathbf{x}^i \rangle \geq 1 - \xi_i \equiv 0 \geq 1 - \xi_i \text{ - } y^i \cdot \langle \mathbf{w}, \mathbf{x}^i \rangle \quad \forall i \in [\text{n}]$$

$$\text{Let } \text{g}_i(\xi_i) = 1 - \xi_i - y^i \cdot \langle \mathbf{w}, \mathbf{x}^i \rangle \quad \forall i \in [\text{n}]$$

So the problem P2 can be redefined as:

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d} \{\tfrac{1}{2}||\mathbf{w}||_2^2 \ + \ C\sum_{i=1}^n \xi_i^2\}$$

$$\text{st. } \text{g}_i(\xi_i) \leq 0 \quad \forall i \in [\text{n}]$$

The Lagrangian for the problem is

$$\mathbf{L(w,\boldsymbol{\xi},\boldsymbol{\alpha})} = \mathbf{\tfrac{1}{2}||w||_2^2} \ + \ \boldsymbol{C\sum_{i=1}^n \xi_i^2} \ + \ \boldsymbol{\sum_{i=1}^n \alpha_i \cdot g_i(\xi_i)}$$

This transforms our P2 to

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d \ \boldsymbol{\xi}\in\mathbb{R}^n} \{ \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \mathbf{L}(\mathbf{w},\xi,\alpha) \} \text{ st. } \alpha_i \geq 0 \quad \forall i \in [\text{n}]$$

2 Part 2

The problem P2 was redefined as:

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d \ \boldsymbol{\xi}\in\mathbb{R}^n} \{ \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \{ \ \tfrac{1}{2}||\mathbf{w}||_2^2 \ + \ C\sum_{i=1}^n \xi_i^2 \ + \ \sum_{i=1}^n \alpha_i \cdot g_i(\xi_i) \ \} \ \}$$

$$\text{st. } \alpha_i \geq 0, \forall i \in [\text{n}] \text{ (g}(\xi) \text{ as defined in part 1)}$$

The corresponding dual problem is:

$$\arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \{ \arg\min_{\mathbf{w}\in\mathbb{R}^d \ \xi\in\mathbb{R}^n} \{ \ \tfrac{1}{2}||\mathbf{w}||_2^2 \ + \ C\sum_{i=1}^n \xi_i^2 \ + \ \sum_{i=1}^n \alpha_i \cdot g_i(\xi_i) \ \}$$

$$\text{st. } \alpha_i \geq 0 \quad \forall i \in [\text{n}]$$

Preprint. Under review.

Applying the first order optimality trick in $\underset{\mathbf{w}\in\mathbb{R}^d\ \xi\in\mathbb{R}^n}{\arg\min}\ \{\ \frac{1}{2}||\mathbf{w}||_2^2\ +\ C\sum_{i=1}^n \xi_i^2\ +\ \sum_{i=1}^n \alpha_i\cdot(1-\xi_i - y^i\langle\mathbf{w},\mathbf{x}^i\rangle)\}$ to eliminate **w** and $\boldsymbol{\xi}$ $\frac{\partial\mathbf{L}}{\partial\mathbf{w}}=0$ and $\frac{\partial\mathbf{L}}{\partial\boldsymbol{\xi}}=0$:

$$\nabla_{\mathbf{w}}\mathbf{L} = [\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, .., \frac{\partial L}{\partial w_d}]$$

$$\implies \nabla_{\mathbf{w}}\mathbf{L} = [w_1 - \sum_{i=1}^n y^i x_1^i \alpha_i, w_2 - \sum_{i=1}^n y^i x_2^i \alpha_i, ..., w_n - \sum_{i=1}^n y^i x_n^i \alpha_i]$$

$$\implies \nabla_{\mathbf{w}}\mathbf{L} = \mathbf{w} - \sum_{i=1}^n \alpha_i \cdot y^i \cdot \mathbf{x}^i$$

$$\text{At Extrema } \nabla_{\mathbf{w}}\mathbf{L} = 0 \implies \mathbf{w} = \sum_{i=1}^n \alpha_i \cdot y^i \cdot \mathbf{x}^i$$

And since $\mathbf{H_w}(\mathbf{L}) = \mathbf{I}$ (at $\mathbf{w} = \sum_{i=1}^n \alpha_i \cdot y^i \cdot \mathbf{x}^i$) which is a PSD so the extrema in this case is a minima (it is actually a global minima since there is only one minima of this function) .

Also,

$$\nabla_{\boldsymbol{\xi}}\mathbf{L} = [\frac{\partial L}{\partial \xi_1}, \frac{\partial L}{\partial \xi_2}, .., \frac{\partial L}{\partial \xi_n}]$$

$$\implies \nabla_{\boldsymbol{\xi}}\mathbf{L} = [2C\xi_1 - \alpha 1, 2C\xi_2 - \alpha 2, ..., 2C\xi_n - \alpha_n]$$

$$\implies \nabla_{\boldsymbol{\xi}}\mathbf{L} = 2C\boldsymbol{\xi} - \boldsymbol{\alpha}$$

$$\text{At Extrema } \nabla_{\boldsymbol{\xi}}\mathbf{L} = 0 \implies \boldsymbol{\xi} = \frac{1}{2C}\boldsymbol{\alpha}$$

clearly $\mathbf{H_\xi}(\mathbf{L}) = 2C\mathbf{I}$ (at $\boldsymbol{\xi} = \frac{1}{2C}\boldsymbol{\alpha}$) which is a PSD (since C $\geq$ 0) so the extrema in this case is a minima (it is actually a global minima since there is only one minima of this function)

Substituting **w** and $\boldsymbol{\xi}$ in the dual problem

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^n}{\arg\max}\ \{\ \underset{\mathbf{w}\in\mathbb{R}^d\ \xi\in\mathbb{R}^n}{\arg\min}\ \{\ \frac{1}{2}||\mathbf{w}||_2^2\ +\ C\sum_{i=1}^n \xi_i^2\ +\ \sum_{i=1}^n \alpha_i\cdot g_i(\xi_i)\ \}$$

$$\text{st. } \alpha_i \geq 0\ \ \forall i \in \text{[n]}$$

$$\equiv \underset{\alpha\in\mathbb{R}^n}{\arg\max}\{\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y^i y^j \langle\mathbf{x}^i,\mathbf{x}^j\rangle\ +\ \sum_{i=1}^n \frac{\alpha_i^2}{4C}\ +\ \sum_{i=1}^n \alpha_i\cdot(1-\frac{\alpha_i}{2C} - \langle\sum_{j=1}^n \alpha_j y^j \mathbf{x}^j, \mathbf{x}^i\rangle\cdot y^i)\}$$

$$\equiv \underset{\alpha\in\mathbb{R}^n}{\arg\max}\{\sum_{i=1}^n(\alpha_i - \frac{\alpha_i^2}{4C}) - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y^i y^j\langle\mathbf{x}^i,\mathbf{x}^j\rangle\ \}$$

$$\text{st. } \alpha_i \geq 0, \forall i \in \text{[n]}$$

So the final dual problem after eliminating **w** and $\boldsymbol{\xi}$ is

$$\underset{\alpha\in\mathbb{R}^n}{\arg\max}\{\sum_{i=1}^n(\alpha_i - \frac{\alpha_i^2}{4C}) - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y^i y^j\langle\mathbf{x}^i,\mathbf{x}^j\rangle\ \}$$

$$\text{st. } \alpha_i \geq 0, \forall i \in \text{[n]}$$

3 Part 3

Method 1: Mini Batch SGD on P1

Firstly we hid the bias term in w and add a column of 1 in X

Gradient expression in this method is:

$$gradient = \mathbf{w} + C.\frac{n}{B}\sum_{i=s_1}^{s_B}\partial(l_i(\mathbf{w}))$$

$$\text{where } l_i(\mathbf{w}) = ([1 - y^i\langle \mathbf{w}, \mathbf{x}^i\rangle]_+)^2$$

$$\partial(l_i(\mathbf{w})) \;=\; \{ \quad \begin{matrix} 0 & (1 - y^i \cdot \langle \mathbf{w}, \mathbf{x}^i\rangle) \leq 0 \\ -2y^i(1 - y^i\langle \mathbf{w}, \mathbf{x}^i\rangle)\mathbf{x}^i & (1 - y^i \cdot \langle \mathbf{w}, \mathbf{x}^i\rangle) > 0 \end{matrix}$$

And to update w with each iteration we used

$$\mathbf{w} = \mathbf{w} - \eta * gradient$$

where B is the batch size and $\{s_1, s_2, ..., s_B\}$ are the indices of the elements in the batch.

We used Batch size of $B = 500$ and an initial step length $\eta = 0.000003$ which gave us the best results. Also, changing them a bit didn't affect the minimum objective value of P1 achieved by the algorithm.

Method 2: Coordinate Descent on P1

Partial derivative for c^{th} coordinate in this method is defined as:

$$\nabla_c L(\mathbf{w}) = w_c + C.\sum_{i=1}^{n} CoordSubGrad(c, y^i, \mathbf{w}, \mathbf{x}^i)$$

here $L(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}([1 - y^i\langle \mathbf{w}, \mathbf{x}^i\rangle]_+)^2$ and $CoordSubGrad(c, y^i, \mathbf{w}, \mathbf{x}^i)$ is :

$$CoordSubGrad(c, y^i, \mathbf{w}, \mathbf{x}^i) \;=\; \{ \quad \begin{matrix} 0 & (1 - y^i \cdot \langle \mathbf{w}, \mathbf{x}^i\rangle) \leq 0 \\ -2y^i x_c^i(1 - y^i\langle \mathbf{w}, \mathbf{x}^i\rangle) & (1 - y^i \cdot \langle \mathbf{w}, \mathbf{x}^i\rangle) > 0 \end{matrix}$$

Updation of c^{th} coordinate in each iteration will be done as follows:

$$w_c = w_c - \eta * \nabla_c L(\mathbf{w})$$

We used cyclic method on a random permutations in this method. We created a random permutation and used cyclic method on it to update the coordinates. When we reach the end of the permutation we create a new random permutation and repeat the above method.
We chose step length $\eta = 0.00009$ and kept the step length η constant in each iteration. This gave us the best result.

Method 3: Coordinate Maximization on D2

By applying the first order optimality on dual problem the formula obtained for updating any dual variable α_i is:

$$\alpha_i = \Pi(\tfrac{2C(1-p)}{2Cq+1})$$

$$\Pi(x) \;=\; \{ \quad \begin{matrix} 0 & x \leq 0 \\ x & x > 0 \end{matrix}$$

C, p,and q is defined as follows:

C is the hyper parameter present in the objective function

$$p = y^i \sum_{j \neq i} \alpha_j y^j \langle \mathbf{x}^i, \mathbf{x}^j \rangle$$

$$q = \sum_{i=1}^{n} \|\mathbf{x}^i\|^2$$

$\alpha_{i's}$ were initialized to 0 and the value of q is calculated initially which can be used for all iterations. For choosing any coordinate first we create a random permutation of numbers 1 to n where n is number of data points. Then we iterate over that permutation and update the α_i according to the element of the permutation. When we reach the end of the permutation, we create a new random permutation and repeat the above method.
As the calculation of p in each iteration takes $O(nd)$ time where n is number of the data points and d is the dimension of the data points, we use the value of **w** obtained while calculating the dual problem to optimize the calculation of p.

$\mathbf{w} = \sum_{i=1}^{n} \alpha_i \cdot y^i \cdot \mathbf{x}^i$

Simplifying the expression of p we get $\implies \quad p = y^i(\langle w, \mathbf{x}^i \rangle - \alpha_i \cdot y^i q)$

So, to calculate p we update w in each iteration. So, using new expression of p it can be calculated in $O(d)$ in each iteration.

4 Part 4

Method 1: Mini Batch SGD on P1

First we used the step length η at time t to be $\eta/\sqrt{t}$ and used the values of η as 0.1, 0.01, 0.001, 0.0001, 0.00001 with Batch size of $B = 5$ among which $\eta = 0.0001$ gave best results on P1 objective value. But then we changed used step length at any time t to be η on values of η as 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001, 0.000003 among which $\eta = 0.000003$ gave best results even upon varying Batch size $B = 5, 50, 500$ but $B = 500$ gave the minimum objective value on P1(not much difference though)

Method 2: Coordinate Descent on P1

First we were choosing the random coordinate in each method which gave good results sometimes but many times was not performing well. Then we used the cyclic method to choose the coordinate. It performed better than the previous method. Then we created a random permutation in each iteration and used the cyclic method to choose the coordinate from this permutation. When we reach the end of the permutation then we create a new permutation and repeat the above method. This gave us the best result, i.e. gave the minimum P1 objective value
First we used step length η at time t equal to $\eta/\sqrt{t}$ and then compared it with keeping η constant in each iteration and found that keeping η constant gave better result. We tried different values of η as $0.00001, 0.00003, 0.00005, 0.00007, 0.00009$ and got the best value at $\eta = 0.00009$ (initialisation).

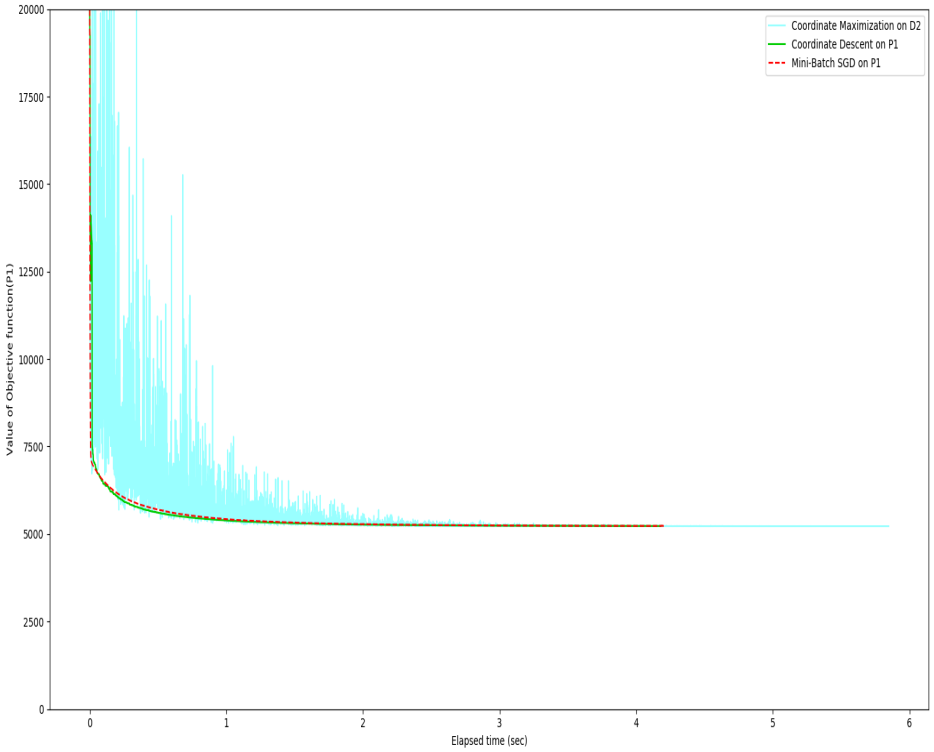
Method 3: Coordinate Maximization on D2

We initialized $\boldsymbol{\alpha}$ with 0 (reason mentioned below).
First we tried the method which chooses the random coordinate in each iteration but it didn't give good results a many times. Then we tried the cyclic method which gave better results than random method. Then we tried the cyclic method over random permutations. We created a random permutation and applied cyclic method on it and when we reached the end of the permutation we created a new random permutation and repeated the above method. The cyclic method using random permutations gave the best results.

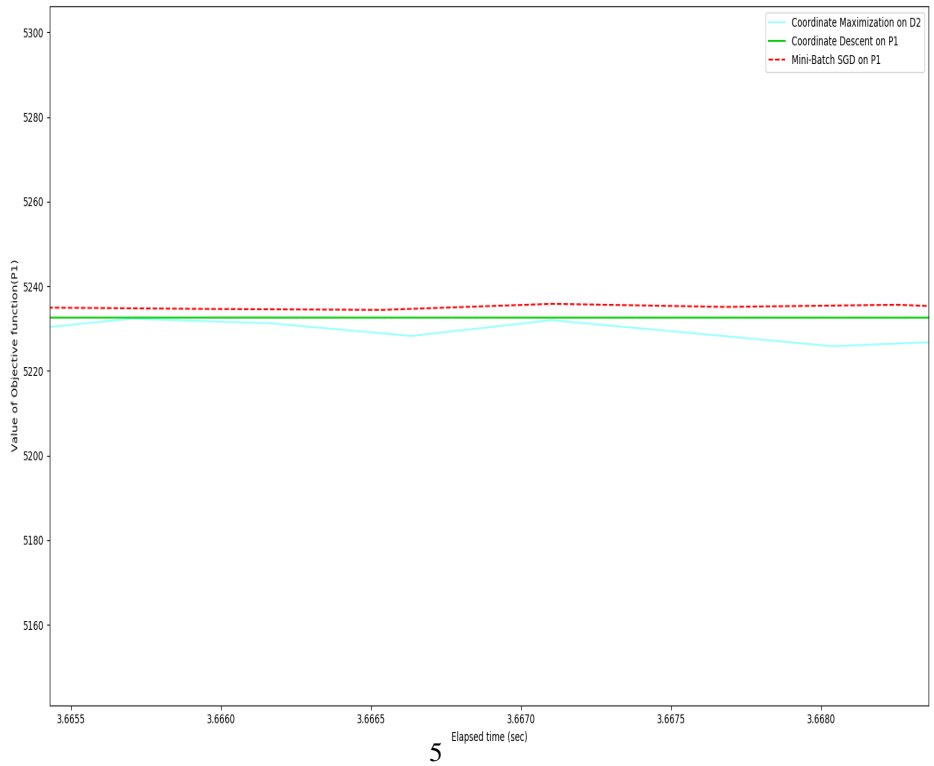
Tried each on different initialisation of α (i.e. all ones, all zeros, all Cs) where zero vector gave the best results (i.e. minimum objective value)

5 Part 5

The convergence curves obtained by the above 3 methods are:



The zoomed image of the above graph in the region where each curve converges is below:



We think coordinate maximisation on Dual performs the best as comparing the minimum values obtained by all three, co-ordinate maximisation gives the least value which becomes constant after a while with almost negligible deviation.

6 Part 7

Consider P2 with the positivity constraint

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min}\ \tfrac{1}{2}||\mathbf{w}||_2^2\ +\ C\textstyle\sum_{i=1}^n \xi_i^2$$

$$\text{st. } y^i\cdot\langle\,\mathbf{w},\mathbf{x}^i\rangle \geq 1-\xi_i\ \ and\ \xi_i\geq 0\ \equiv\ 1-\xi_i-y^i\cdot\langle\,w,\mathbf{x}^i\rangle\ \leq 0\ and\ \ -\xi_i\leq 0\quad \forall i\in[n]$$

Removing constraints, the Lagrangian for the problem comes out to be

$$\mathrm{L}(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\alpha})=\frac{1}{2}||\mathbf{w}||_2^2\ +\ C\sum_{i=1}^n\xi_i^2\ +\ \sum_{i=1}^n\alpha_i\cdot(1-\xi_i-y^i\cdot\langle\,w,\mathbf{x}^i\rangle)+\sum_{i=1}^n\beta_i\cdot(-\xi_i)$$
$$\alpha_i\geq 0\ and\ \beta_i\geq 0\quad\forall i\in[n]$$

Using the Lagrangian, P2 transforms to

$$\underset{\mathbf{w}\in\mathbb{R}^d,\boldsymbol{\xi}\in\mathbb{R}^n}{\arg\min}\ \{\ \underset{\boldsymbol{\beta},\boldsymbol{\alpha}\in\mathbb{R}^n}{\arg\max}\{\ \frac{1}{2}||\mathbf{w}||_2^2\ +\ C\sum_{i=1}^n\xi_i^2+\sum_{i=1}^n\beta_i\cdot(-\xi_i)+\sum_{i=1}^n\alpha_i\cdot(1-\xi_i-y^i\cdot\langle\,w,\mathbf{x}^i\rangle)\ \}\ \}$$
$$st.\ \alpha_i\geq 0\ and\ \beta_i\geq 0\quad\forall i\in[n]$$

The corresponding dual problem is

$$\underset{\boldsymbol{\beta},\boldsymbol{\alpha}\in\mathbb{R}^n}{\arg\max}\{\ \underset{\mathbf{w}\in\mathbb{R}^d,\boldsymbol{\xi}\in\mathbb{R}^n}{\arg\min}\ \{\frac{1}{2}||\mathbf{w}||_2^2\ +\ C\sum_{i=1}^n\xi_i^2+\sum_{i=1}^n\beta_i\cdot(-\xi_i)+\sum_{i=1}^n\alpha_i\cdot(1-\xi_i-y^i\cdot\langle\,w,\mathbf{x}^i\rangle)\}\}$$
$$st.\ \alpha_i\geq 0\ and\ \beta_i\geq 0\quad\forall i\in[n]---------- (1)$$

We know P2 is a convex function in terms of **w** and $\boldsymbol{\xi}$ (since C $\geq$ 0), hence there is a strong duality between P2 and its Dual problem which implies that the **w** and $\boldsymbol{\xi}$ of P2 and the Dual problem at the optimum point will be same. Consider

$$\mathbf{L}=\frac{1}{2}||\mathbf{w}||_2^2\ +\ C\sum_{i=1}^n\xi_i^2+\sum_{i=1}^n\beta_i\cdot(-\xi_i)+\sum_{i=1}^n\alpha_i\cdot(1-\xi_i-y^i\cdot\langle\,w,\mathbf{x}^i\rangle)$$

We need

$$\underset{\boldsymbol{\beta},\boldsymbol{\alpha}\in\mathbb{R}^n}{\arg\max}\{\ \underset{\mathbf{w}\in\mathbb{R}^d,\boldsymbol{\xi}\in\mathbb{R}^n}{\arg\min}\ \{\mathbf{L}\}\}st.\ \alpha_i\geq 0\ and\ \beta_i\geq 0\quad\forall i\in[n]$$

Applying the first order optimality condition

$$\nabla_{\mathbf{w}}\mathbf{L}=\mathbf{w}-\sum_{i=1}^n\alpha_i\cdot y^i\mathbf{x}^i$$

At Extrema $\nabla_{\mathbf{w}}\mathbf{L}=0\implies \mathbf{w}=\sum_{i=1}^n\alpha_i\cdot y^i\cdot\mathbf{x}^i$

and $\mathbf{H}_{\mathbf{w}}(\mathbf{L})=\mathbf{I}$, which is a PSD implies a minima

$$\nabla_{\boldsymbol{\xi}}\mathbf{L}=2C\boldsymbol{\xi}-(\boldsymbol{\alpha}+\boldsymbol{\beta})$$

At Extrema $\nabla_{\boldsymbol{\xi}}\mathbf{L}=0\implies \boldsymbol{\xi}=\frac{1}{2C}(\boldsymbol{\alpha}+\boldsymbol{\beta})$

and $\mathbf{H}_{\boldsymbol{\xi}}(\mathbf{L})=2C\mathbf{I}$, which is a PSD implies a minima

Substituting **w** and $\boldsymbol{\xi}$ gives us

$$\underset{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^n}{\arg\max}\{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j\rangle \ + \ \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\frac{(\beta_i+\alpha_i)^2}{4C}\}$$

$$\text{st. } \alpha_i \geq 0 \ \ \beta_i \geq 0 \quad \forall i \in \text{[n]}$$

Let E = $\{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j\rangle \ + \ \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\frac{(\beta_i+\alpha_i)^2}{4C}\}$

And let the optimal solution for E occurs at $\{\alpha_1, \alpha_2, ..., \alpha_n, \beta 1, \beta_2, .., \beta_n\}$ let the value attained at the given point be e under the constraint mentioned above.

Claim:Let value of E at point $\{\alpha_1, \alpha_2, ..., \alpha_n, 0, 0, .., 0\}$ be e' then $e' \geq e$

Proof:

$$e' - e = \sum_{i=1}^{n}\frac{(\beta_i+\alpha_i)^2}{4C} - \sum_{i=1}^{n}\frac{(0+\alpha_i)^2}{4C} = \sum_{i=1}^{n}\frac{\alpha_i^2 + 2\beta_i\alpha_i}{4C} \geq 0$$

$$as \ \alpha_i, \beta_i \geq 0 \ \forall i \in [n]$$

$$\implies e' \geq e$$

So in the maximum value of expression E (maxima) all the β_i's will be 0 (otherwise it can't be maxima as proved by claim above). ————(2)

Clearly the above problem can be reduced to (since all the β_i's are 0 in the optimal solution):

$$\underset{\alpha\in\mathbb{R}^n}{\arg\max}\{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j\rangle \ + \ \sum_{i=1}^{n}(\frac{-\alpha_i^2}{4C} + \alpha_i)$$

$$\text{st. } \alpha_i \geq 0 \quad \forall i \in \text{[n]}$$

So using (2) and substituting it in (1) we get the problem as :

$$\underset{\mathbf{b},\boldsymbol{\alpha}\in\mathbb{R}^n}{\arg\min}\{ \ \underset{\mathbf{w}\in\mathbb{R}^d,\boldsymbol{\xi}\in\mathbb{R}^n}{\arg\max} \ \{\frac{1}{2}||\mathbf{w}||_2^2 \ + \ C\sum_{i=1}^{n}\xi_i^2 + \sum_{i=1}^{n}\alpha_i \cdot (1 - \xi_i - y^i \cdot \langle w, \mathbf{x}^i\rangle)\}\}$$

$$\text{st. } \alpha_i \geq 0 \quad \forall i \in [n]$$

which is equivalent to the problem:

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} \ \frac{1}{2}||\mathbf{w}||_2^2 \ + \ C\sum_{i=1}^{n}\xi_i^2$$

$$\text{st. } y^i \cdot \langle w, \mathbf{x}^i\rangle \geq 1 - \xi_i \quad \forall i \in \text{[n]}$$

Which is P2 without positivity constraint. Hence, both have same solution.