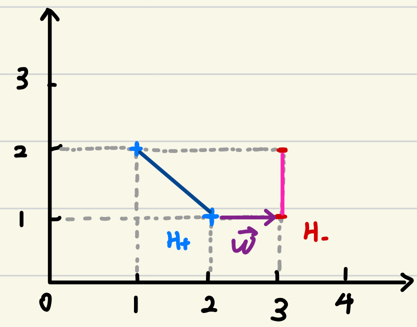
Q1:

$$\text{margin} = \min_{\substack{u \in \Delta, \\ v \in \Delta}} \frac{1}{2} \Big\| \sum_{i \in H_+} u_i x_i - \sum_{i \in H_-} v_j x_j \Big\|_2^2$$



By inspection, the minimum distance between these two hyperplanes H_+ and H_- is the distance between (2,1) and (3,1)

Let r be a scale such that
$$\vec{w} = r\left(\begin{bmatrix}2\\1\end{bmatrix} - \begin{bmatrix}3\\1\end{bmatrix}\right) = r\begin{bmatrix}-1\\0\end{bmatrix} = \begin{bmatrix}-r\\0\end{bmatrix}$$

Since $\vec{w}^T x + b = y$, and $x_1 = \begin{bmatrix}2\\1\end{bmatrix}$, $y_1 = 1$

$x_2 = \begin{bmatrix}3\\1\end{bmatrix}$, $y_2 = -1$,

$$\begin{cases} \vec{w}^T x_1 + b = y_1 \\ \vec{w}^T x_2 + b = y_2 \end{cases}$$

$$\Rightarrow \begin{cases} [-r \quad 0]\begin{bmatrix}2\\1\end{bmatrix} + b = 1 \\ [-r \quad 0]\begin{bmatrix}3\\1\end{bmatrix} + b = -1 \end{cases} \Rightarrow \begin{cases} -2r + b = 1 & ① \\ -3r + b = -1 & ② \end{cases}$$

$\Rightarrow$ ① − ② : $r = 2$.

Sub. $r = 2$ into ① : $b = 1 + 2r = 5$

$$\Rightarrow \begin{cases} r = 2 \\ b = 5 \end{cases}$$

Therefore, $w^* = \begin{bmatrix}-r\\0\end{bmatrix} = \begin{bmatrix}-2\\0\end{bmatrix}$ and $b^* = 5$ is the hard-margin SVM optimum solution.

Q2 :

$$\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \ \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\max\{1-y_i(w^T x_i+b),0\}^2$$

This is equivalent to
$$\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R},\, \varepsilon\in\mathbb{R}^n} \ \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\varepsilon_i^2$$

$$\text{s.t. } \forall i,\ (1-y_i\hat{y}_i)_+^2 = \max\{1-y_i\hat{y}_i,0\}^2 \le \varepsilon_i^2$$

$$\Rightarrow \begin{cases}(1-y_i\hat{y}_i)^2 \le \varepsilon_i^2 \\ 0 \le \varepsilon_i^2\end{cases}$$

$$\hat{y}_i = w^T x_i + b$$

Since $\varepsilon_i^2 \ge 0$ for every ε_i, we don't need the second condition

Since $(1-y_i\hat{y}_i)^2 \le \varepsilon_i^2$, $1-y_i\hat{y}_i \le \varepsilon_i$

The problem is converted to:
$$\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R},\, \varepsilon\in\mathbb{R}^n} \ \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\varepsilon_i^2$$

$$\text{s.t. } \forall i \quad 1-y_i\hat{y}_i \le \varepsilon_i$$

Lagrangian is : $L(w,b,\varepsilon,\alpha) = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\varepsilon_i^2 + \sum_{i=1}^{n}\alpha_i(1-y_i(w^T x_i+b)-\varepsilon_i)$

where $\alpha \ge 0$

Lagrangian Dual : $\min_{w,b,\varepsilon}\ \max_{\alpha\ge 0,\, \beta\le 0} L(w,b,\varepsilon,\alpha,\beta)$

$$\Rightarrow \max_{\alpha\ge 0}\ \min_{w,b,\varepsilon} L(w,b,\varepsilon,\alpha)$$

$$\Rightarrow \max_{\alpha\ge 0}\ \min_{w,b,\varepsilon} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\varepsilon_i^2 + \sum_{i=1}^{n}\alpha_i(1-y_i(w^T x_i+b)-\varepsilon_i)$$

Gradients : $$\frac{\partial L}{\partial w} = \frac{1}{2}\frac{\partial(\|w\|_2^2)}{\partial w} + \sum_{i=1}^{n}\frac{\partial(-\alpha_i y_i w^T x_i)}{\partial w}$$

$$= \frac{1}{2}\cdot 2w + \sum_{i=1}^{n}(-\alpha_i y_i x_i)$$

$$= w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0$$

$$\Rightarrow w^* = \sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \frac{\partial(\alpha_i y_i b)}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \varepsilon_i} = C \frac{\partial(\varepsilon_i^2)}{\partial \varepsilon_i} + \frac{\partial(-\alpha_i \varepsilon_i)}{\partial \varepsilon_i}$$

$$= 2C\varepsilon_i - \alpha_i = 0$$

$$\Rightarrow \varepsilon_i^* = \frac{\alpha_i}{2C} \geq 0 \quad \text{as} \quad \alpha_i \geq 0$$

Sub. into L:

$$\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \varepsilon_i^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i(w^T x_i + b) - \varepsilon_i)$$

$$= \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \varepsilon_i^2 + \sum_{i=1}^{n} \alpha_i - \underbrace{\boxed{\sum_{i=1}^{n} \alpha_i y_i x_i}}_{w} w^T - \underbrace{\boxed{\sum_{i=1}^{n} \alpha_i y_i}}_{0} b$$
$$- \sum_{i=1}^{n} \alpha_i \varepsilon_i$$

$$= \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \varepsilon_i^2 + \sum_{i=1}^{n} \alpha_i - ww^T - \sum_{i=1}^{n} \alpha_i \varepsilon_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \underline{\varepsilon_i}^2 - \sum_{i=1}^{n} \alpha_i \underline{\varepsilon_i}$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \left(\frac{\alpha_i}{2C}\right)^2 - \sum_{i=1}^{n} \alpha_i \frac{\alpha_i}{2C}$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\|w\|_2^2 + C \cdot \frac{1}{4C^2} \sum_{i=1}^{n} \alpha_i^2 - \frac{1}{2C}\sum_{i=1}^{n} \alpha_i^2$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\|w\|_2^2 + \frac{1}{4C}\sum_{i=1}^{n} \alpha_i^2 - \frac{1}{2C}\sum_{i=1}^{n} \alpha_i^2$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\|\underline{w}\|_2^2 - \frac{1}{4C}\sum_{i=1}^{n} \alpha_i^2$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\left\|\sum_{i=1}^{n} \alpha_i y_i x_i\right\|_2^2 - \frac{1}{4C}\sum_{i=1}^{n} \alpha_i^2$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j - \frac{1}{4C}\sum_{i=1}^{n} \alpha_i^2$$

The dual is:

$$\max_{\alpha \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j - \frac{1}{4C} \sum_{i=1}^{n} \alpha_i^2$$

such that $\sum_{i=1}^{n} \alpha_i y_i = 0$,

which is equivalent to:

$$\min_{\alpha \geq 0} -\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j + \frac{1}{4C} \sum_{i=1}^{n} \alpha_i^2$$

such that $\sum_{i=1}^{n} \alpha_i y_i = 0$.

Q3: $c\,(1-y_i\hat{y_i})_+^2 \qquad \hat{y_i} = w^T X_i + b \qquad (2)$

Case 1: If $1-y_i\hat{y_i} \le 0$, $\quad (1-y_i\hat{y_i})_+ = 0$

$$c\,(1-y_i\hat{y_i})_+^2 = 0$$

$$\frac{\partial}{\partial w} = 0$$

$$\frac{\partial}{\partial b} = 0$$

Case 2: If $1-y_i\hat{y_i} > 0$, $\quad (1-y_i\hat{y_i})_+ = 1-y_i\hat{y_i}$

$$c\,(1-y_i\hat{y_i})_+^2 = c\,(1-y_i\hat{y_i})^2$$

$$= c\,(1-y_i(w^TX_i - b))^2$$

$$= c\,(1-y_iX_iw^T - y_ib)^2$$

$$\frac{\partial}{\partial w} = c\cdot 2(1-y_iX_iw^T - y_ib)(-y_iX_i)$$

$$= -2c\,y_iX_i\,(1-y_i\hat{y_i}) \quad \text{as} \quad 1-y_iX_iw^T - y_ib = 1-y_i\hat{y_i}$$

$$\frac{\partial}{\partial b} = c\cdot 2(1-y_iX_iw^T - y_ib)(-y_i)$$

$$= -2c\,y_i\,(1-y_iX_iw^T - y_ib)$$

$$= -2c\,y_i\,(1-y_i\hat{y_i}) \quad \text{as} \quad 1-y_iX_iw^T - y_ib = 1-y_i\hat{y_i}$$

Therefore, by combining the two cases,
the gradient of equation (2) with respect to w is $\underline{-2c\,y_iX_i\,(1-y_i\hat{y_i})_+}$,
and the gradient of equation (2) with respect to b is $\underline{-2c\,y_i\,(1-y_i\hat{y_i})_+}$

Q4 :

$$P^{\eta}(w) = \arg\min_{z} \frac{1}{2\eta} \|z-w\|_2^2 + \frac{1}{2} \|z\|_2^2$$

$$\frac{\partial}{\partial z} = \frac{1}{2\eta} \frac{\partial}{\partial z} (\|z-w\|_2^2) + \frac{1}{2} \frac{\partial}{\partial z} (\|z\|_2^2)$$

$$= \frac{1}{2\eta} \cdot 2(z-w) + \frac{1}{2} \cdot 2z$$

$$= \frac{1}{\eta} (z-w) + z$$

Let $\frac{\partial}{\partial z} = 0$.

$$\frac{1}{\eta} (z-w) + z = 0$$

$$\frac{1}{\eta} z - \frac{1}{\eta} w + z = 0$$

$$\frac{1+\eta}{\eta} z = \frac{1}{\eta} w$$

$$z = \frac{1}{\eta} w \cdot \frac{\eta}{1+\eta}$$

$$z = \frac{1}{1+\eta} w$$

Q5 :

From Q3, $\frac{\partial C(1-y_i\hat{y}_i)_+}{\partial w} = -2C\, y_i x_i\, (1-y_i\hat{y}_i)_+$

$$\frac{\partial C(1-y_i\hat{y}_i)_+}{\partial b} = -2C\, y_i\, (1-y_i\hat{y}_i)_+$$

Equation (1) can be converted to:

$$\min_{w,b}\ \underline{\tfrac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}(1-y_i\hat{y}_i)_+^2} \qquad \hat{y}_i = w^T x_i + b$$

$$\frac{\partial}{\partial w} = \frac{1}{2}\cdot 2w + (-2C)\sum_{i=1}^{n}(1-y_i\hat{y}_{i+})$$

$$= w - 2C\sum_{i=1}^{n}(1-y_i\hat{y}_i)$$

$$\frac{\partial}{\partial b} = -2C\sum_{i=1}^{n} y_i(1-y_i\hat{y}_i)$$

From Q4, $z = \frac{1}{1+\eta} w$

Algorithm 1: SGD for SVM.

Input: $X \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \{-1,1\}^n$, $\mathbf{w} = \mathbf{0}_d$, $b = 0$, max_pass $\in \mathbb{N}$, step size η
Output: $\mathbf{w}, b$

```
for t = 1, 2, ..., max_pass do
    for i = 1, 2, ..., n do
        choose step size η
        if y_i(⟨x_i, w⟩ + b) ≤ 1 then
            w_t ← w_{t-1} − η ∇w_{t-1}                  // x_i is the i-th row of X
            b_t ← b_{t-1} − η ∇b_{t-1}
        w_t ← (1/(1+η)) w_t                             // the proximal step to maximize margin
```

The implementation is in "code_part.ipynb".

Q6:

After running the dataset in "code-part.ipynb", the result is :

```
C
10000

w
array([-1.99805291e+00, -7.34489503e-04])

b
4.995740369998062
```

$w = \begin{bmatrix} -1.998 \\ -7.344\times 10^{-4} \end{bmatrix}$ is very closed to $w^* = \begin{bmatrix} -2 \\ 0 \end{bmatrix}$

$b = 4.99574$ is very closed to $b^* = 5$.

Hence, by using a large C, which is 10000, I recover the hard-margin SVM solution in Q1 be