Q2)

Optimization problem:

$$\min_{\omega, b, \varepsilon} \frac{1}{2}\|\omega\|^2 + \frac{c}{2}\sum_{i=1}^{m} \varepsilon^2$$

$$\text{s.t.} \quad y^{(i)}(\omega^T x^{(i)} + b) \geq 1 - \varepsilon_i \quad (i \in \{1, \dots, m\})$$

(a) There is no need for the non-negativity constraint on ε, i.e. $\varepsilon \geq 0$ for l_2 norm soft margin optimization

~~If~~

In the equation $y^{(i)}(\omega^T x^{(i)} + b) \geq 1 - \varepsilon_i$, ε_i is introduced only to consider some of the wrongly classified samples so as to make the model more generalisable.

ε_i can be visualised as the distance or the slack given to the misclassification. It is the deviation ~~for~~ from the ~~int~~ pattern or ~~the~~ so to say, 'the glitch in the matrix'

Now the sign of this variable is defined by the relative position of the misclassifications from the ~~SV plane~~ SV hyperplane. On one side it is +ve while on the other it is -ve.

Now in the equation for l_2 norm optimization, the term is squared.

$\therefore$ This eliminates the sign of ε as a result eliminating the need of keeping a constraint on it's parity.

(b) The lagrangian for the given problem is:

$$\mathcal{L}(\omega, \alpha, b, \varepsilon) = \frac{1}{2}\omega^T\omega + \frac{c}{2}\sum_{i=1}^{m}\varepsilon_i^2$$

$$- \sum_{i=1}^{m}\alpha_i\left(y^{(i)}(\omega^T x^{(i)} + b) + \varepsilon_i - 1\right)$$

$$\forall \alpha_i \geq 0 \quad \& \quad \forall i \in \{1, ---, m\}$$

(c) For calculating the Dual we need some derivatives first & equate them to 0.

$$\frac{\partial L}{\partial \omega} = \left(\frac{1}{2}\right)(2\omega) + 0 - \sum_{i=1}^{m}(\alpha_i)(y_i x_i) = 0$$

$$\therefore \omega = \sum_{i=1}^{m}(\alpha_i)(y_i x_i).$$

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

$$\Rightarrow \sum_{i=1}^{m}(\alpha_i)(y_i) = 0.$$

$$\frac{\partial L}{\partial \varepsilon} = 0 - \sum_{i=1}^{M}\alpha_i + 2\frac{c}{2}\sum_{i=1}^{m}\varepsilon_i = 0.$$

$$\Rightarrow \sum_{i=1}^{m}\alpha_i = c\sum_{i=1}^{m}\varepsilon_i.$$

Putting the above & in Primal form,

$$\text{Dual}(D) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j(x_i)^T x_j y_i y_j + \frac{1}{2}\sum_{i=1}^{m}\alpha_i\varepsilon_i$$

$$- \sum_{i=1}^{m}\alpha_i\left(y_i\left(\left(\sum_{j=1}^{m}\alpha_j y_j x_j\right)^T x_i + b\right) + \varepsilon_i - 1\right)$$

on simplifying.

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

→ maximize this ~~two~~ over α. , s.t. $\alpha_i \geq 0 \ \forall i \in \{1, \dots, m\}$

and $\sum_{i=1}^{m} \alpha_i y_i = 0$

Q3) Gaussian Kernel : $K(x,z) = \exp(-\|x-z\|^2/\tau^2)$.

Given : no of identical pts in training set

T.P. : we can always find a val. of bandwidth param τ s.t. SVM achieves 0 error.

(a) $f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} K(x^{(i)}, x) + b.$

(there is min sep^n dist ε s.t. $\|x^{(j)} - x^{(i)}\| \geq \varepsilon$

To find: set of params $\{\alpha_1, \ldots, \alpha_m, b\}$ & Gaussian Kernel width τ s.t. $x^{(i)}$ is classified correctly $\forall i$.

(ans) from hint, let $\alpha_i = 1 \ \forall i$ & $b = 0$

Let any arbitrary traing point be $\{x^{(i)}, y^{(i)}\}$

now prediction would be correct if

$$|f(x^{(i)}) - y^{(i)}| < 1. \quad -(1)$$

$$\therefore |f(x^{(i)}) - y^{(i)}| = \left|\sum_{k=1}^{m} y^{(k)} K(x^{(k)}, x^{(i)}) - y^{(i)}\right|.$$

$$= \left|\sum_{k=1}^{m} y^{(k)} e^{\left(\frac{-\|x^{(k)} - x^{(i)}\|^2}{\tau^2}\right)} - y^{(i)}\right|$$

now from these $k [1, m]$, there is a k which corresponds to i.

$\therefore \left| f(x^{(i)}) - y^{(i)} \right| = \left| \sum_{k \neq i}^{m} y^{(k)} e^{\left(\frac{\| x^{(k)} - x^{(i)} \|^2}{\tau^2} \right)} + \cancel{y^{(i)}} - \cancel{y^{(i)}} \right|$

now from Δ inequality, we know that $|a+b| \leq |a| + |b|$.

If we apply this inequality for each valid k ($m-2$ times total), our eq^n will become:

$$\left| f(x^{(i)}) - y^{(i)} \right| \leq \sum_{k \neq i}^{m} \left| y^{(k)} e^{-\frac{\| x^{(k)} - x^{(i)} \|^2}{\tau^2}} \right|$$

$$\leq \sum_{k \neq i}^{m} |y^{(k)}| \underbrace{\left| e^{-\frac{\| x^{(k)} - x^{(i)} \|^2}{\tau^2}} \right|}_{\text{always +ve.}}$$

now $y^k \in \{-1, 1\}$

$$\therefore \left| f(x^{(i)}) - y^{(i)} \right| \leq \sum_{k \neq i}^{m} e^{-\frac{\| x^{(k)} - x^{(i)} \|^2}{\tau^2}}$$

now $\| x^{(k)} - x^{(i)} \| \geq \varepsilon$ (from assumption given in question)

$$\therefore \left| f(x^{(i)}) - y^{(i)} \right| \leq \sum_{k \neq i}^{m} e^{-\frac{\varepsilon^2}{\tau^2}}$$

since $e^{-\frac{\varepsilon^2}{\tau^2}}$ is not dependent on k.

$$\therefore \left| f(x^{(i)}) - y^{(i)} \right| \leq (m-1) e^{-\frac{\varepsilon^2}{\tau^2}}$$

now from (1)

$$(m-1) e^{-\frac{\varepsilon^2}{\tau^2}} < 1$$

$$\Rightarrow e^{-\frac{\varepsilon^2}{\tau^2}} < \frac{1}{m-1}$$

taking log on both sides

$$-\frac{\varepsilon^2}{\tau^2} < \log\left(\frac{1}{m-1}\right)$$

$$\tau^2 > \frac{-\varepsilon^2}{\log\left(\frac{1}{m-1}\right)}$$

$\therefore$ for SVM to have 0 error & all the pts to be classified correctly,

$$\tau^2 = \frac{-\varepsilon^2}{\log\left(\frac{1}{m-1}\right)}, \text{ when } \alpha_i = 1 \;\&\; b = 0.$$

$\therefore$ parameters found

(b). Yes the SVM will give us 0 training error, if we ~~use~~ run the SVM with slack variables using the parameter τ found earlier. The ~~points~~ value of τ was calculated in such a way that all the points would have been classified correctly. Since all pts were classified correctly, the training error incurred has to be 0. Now if in this scenario slack variables are introduced, the constraint becomes

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \varepsilon_i \quad \forall i.$$

~~from~~ $\cancel{y^{(i)}(w^T x^{(i)} + b) >}$

With suitable selection of parameters, if the situation ~~exists where a lin hyperplane can be created~~ exists, 0 training error can ~~also~~ be obtained.