$\boxed{\text{Initialization}}$

$W_i \in \mathbb{R}^{d_i \times d_{i-1}} \sim \mathcal{N}(0, 1/d_{i-1})$ per coordinate.

$\boxed{\text{Learning rates}}$ $\eta \in \{0.001, 0.01, 0.1\}$

$\boxed{\text{Data Augmentation}}$

Besides original data, generate new data,

with some computation

$\boxed{\text{HW 2}}$

1. (a) ① $\alpha_i > 0$ then $\max\{\alpha_i, 0\} = \alpha_i$

Then $|\max\{0, \alpha_i\} - \alpha_i| = |\alpha_i - \alpha_i| = 0$

So $|\alpha_i' - \alpha_i| \geq 0$, absolute value ≥ 0. So true.

② $\alpha_i \leq 0$ Then $\max\{\alpha_i, 0\} = 0$

Then $|\max\{0, \alpha_i\} - \alpha_i| = |0 - \alpha_i| = |-\alpha_i| = -\alpha_i$
≥ 0

$\alpha' \in [0, \infty)^n$ so $\alpha_i' \geq 0$ so $\alpha_i' - \alpha_i \geq -\alpha_i$

Since $\alpha_i' - \alpha_i \geq 0$ so $\alpha_i' - \alpha_i = |\alpha_i' - \alpha_i|$

so $|\alpha_i' - \alpha_i| \geq -\alpha_i = |\max\{0, \alpha_i\} - \alpha_i|$

So True.

① Let $\alpha_i > 0$ Then $\max\{\alpha_i, 0\} = \alpha_i$

$\min\{\max\{0, \alpha_i\}, C\} = \min\{\alpha_i, C\} = \alpha_i$

So $|\min\{\max\{0, \alpha_i\}, C\} - \alpha_i| = |\alpha_i - \alpha_i| = 0$

$|\alpha_i' - \alpha_i| \geq 0 = |\min\{\max\{0, \alpha_i\}, C\} - \alpha_i|$

So true

② Let $\alpha_i \leq 0$ Then $\max\{\alpha_i, 0\} = 0$

$\min\{\max\{0, \alpha_i\}, C\} = \min\{0, C\} = 0$

So $|\min\{\max\{0, \alpha_i\}, C\} - \alpha_i| = |0 - \alpha_i| = -\alpha_i$

$\alpha_i' - \alpha_i \geq -\alpha_i \geq 0$ since $\alpha_i' \geq 0$

So $\alpha_i' - \alpha_i = |\alpha_i' - \alpha_i| \geq -\alpha_i =$

$|\min\{\max\{0, \alpha_i\}, C\} - \alpha_i|$

so true.

2. (a) show $\lim_{\sigma \to 0} \frac{f_\sigma(x)}{\exp(-\rho^2/2\sigma^2)} = \sum_{i \in T} \hat{\alpha}_i y_i$

S : support vectors

T : closest support vectors.

Let $f_\sigma(x) = \sum_{i \in S} \hat{\alpha}_i y_i \exp\left(-\frac{\|x - x_i\|_2^2}{2\sigma^2}\right)$

$= \sum_{i \in T} \hat{\alpha}_i y_i \exp\left(-\frac{\|x - x_i\|_2^2}{2\sigma^2}\right)$

$+ \sum_{j \in S/T} \hat{\alpha}_j y_j \exp\left(-\frac{\|x - x_j\|_2^2}{2\sigma^2}\right)$

$$\frac{f_\sigma(x)}{\exp(-\rho^2/2\sigma^2)} = \frac{\sum_{i \in T} \hat{\alpha}_i y_i \left(-\frac{\|x - x_i\|_2^2}{2\sigma^2}\right) + \sum_{j \in S/T} \hat{\alpha}_j y_j \exp\left(-\frac{\|x - x_j\|_2^2}{2\sigma^2}\right)}{\exp(-\rho^2/2\sigma^2)}$$

$$= \frac{\sum_{i \in T} \hat{\alpha}_i y_i \exp\left(-\frac{\rho^2}{2\sigma^2}\right) + \sum_{j \in S/T} \hat{\alpha}_j y_j \exp\left(-\frac{\|x - x_j\|_2^2}{2\sigma^2}\right)}{\exp(-\rho^2/2\sigma^2)}$$

$$= \sum_{i \in T} \hat{\alpha}_i y_i + \sum_{j \in S/T} \hat{\alpha}_j y_j \exp\left(-\frac{\|x - x_j\|_2^2 - \rho^2}{2\sigma^2}\right)$$

$$\lim_{\sigma\to 0} \exp\left(-\frac{\|x-x_j\|_2^2-\rho^2}{2\sigma^2}\right) = \lim_{\sigma\to 0} \exp(-\infty)$$

$$= 0$$

So $$\lim_{\sigma\to 0} \frac{f_\sigma(x)}{\exp(-\rho^2/2\sigma^2)} = \sum_{i\in T} \hat{\alpha}_i y_i$$

(b) Let $k(x_i, x_j) = \exp\left(-\dfrac{\|x_i-x_j\|_2^2}{2\sigma^2}\right)$

So $A_{ij} = y_i y_j k(x_i, x_j)$

$$A\alpha = \begin{bmatrix} \sum_{j=1}^n \alpha_j y_j y_1 k(x_1, x_j) \\ \vdots \\ \sum_{j=1}^n \alpha_j y_j y_n k(x_n, x_j) \end{bmatrix}$$

$$\alpha^T A\alpha = \sum_{i=1}^n \alpha_i \sum_{j=1}^n \alpha_j y_j y_i \, k(x_i, x_j)$$

$$= \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

$$1^T\alpha = \sum_{i=1}^n \alpha_i$$

$$1^T\alpha - \frac{1}{2}\alpha^T A\alpha$$

$$= \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \underbrace{k(x_i, x_j)}_{\exp\left(-\frac{\|x_i-x_j\|_2^2}{2\sigma^2}\right)}$$

(c) Let $\nabla_\alpha h(\alpha) = 0$

So $\nabla_\alpha(-\frac{1}{2}\alpha^T A\alpha) + \nabla_\alpha(1^T\alpha) = 0$

$= A(-\frac{1}{2}\cdot 2\cdot \alpha) + 1^T$

$= -A\alpha + 1^T = 0$

So $A\alpha = 1$

5. (a) $\frac{\partial \hat{R}}{\partial v_j} = \frac{\partial \hat{R}}{\partial f} \cdot \frac{\partial f}{\partial v_j} = l'(f(x; W, v), y) \cdot \sum_{j=1}^{m} \sigma(\langle w_j, x\rangle)$

$$\frac{\partial \hat{R}}{\partial w_j} = \frac{\partial \hat{R}}{\partial f} \cdot \frac{\partial f}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial \langle w_j, x\rangle} \cdot \frac{\partial \langle w_j, x\rangle}{\partial w_j}$$

$$= \frac{\partial \hat{R}}{\partial f} \cdot \frac{\partial f}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial \langle w_j, x\rangle} \cdot x$$

Without loss of generality

(b) Base: $W_p^{(0)} = W_q^{(0)}$

IH: $W_p^{(t-1)} = W_q^{(t-1)} \quad \forall t \geq 1$

$$W_p^{(t)} = W_p^{(t-1)} - \eta \frac{\partial \hat{R}}{\partial W_p^{(t-1)}}$$

Since $W_p^{(t-1)} = W_q^{(t-1)}$

$$\text{So } W_p^{(t-1)} - \eta \frac{\partial \hat{R}}{\partial W_p^{(t-1)}} = W_q^{(t-1)} - \eta \frac{\partial \hat{R}}{\partial W_q^{(t-1)}}$$

So $W_p^{(t)} = W_q^{(t)} \quad \forall t \geq 0$

Same proof applies for $V_p^{(t)} = V_q^{(t)}$

(c) $E\left[\|W^{(0)}x\|_2^2\right]$

$$= E\left[\sum_{j=1}^{m}(w_j^T x)^2\right]$$

$$= \sum_{j=1}^{m} E\left[(w_j^T x)^2\right]$$

$$= x^T x \sum_{j=1}^{m} \underbrace{E\left[(w_j^T)^2\right]}_{Var(w_j^T) - E^2[w_j]}$$

$$= x^T x \sum_{j=1}^{m} \frac{1}{m}$$

$$= x^T x = \|x\|_2^2$$

$$\begin{bmatrix} w_{11} & \cdots & w_{1d} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{md} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} w_1 \cdot x \\ \vdots \\ w_m \cdot x \end{bmatrix}$$