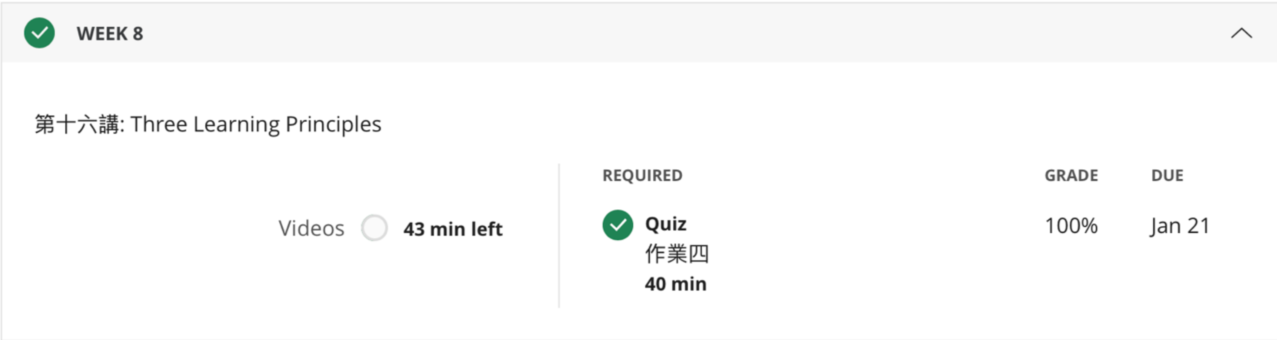
1.




2.

$\because\ \nabla E_{aug}\big(W(t)\big) = \nabla E_{in}\big(W(t)\big) + \frac{2\lambda}{N}\mathrm{W}$

$\because\ \mathrm{W(t+1)} = \mathrm{W(t)} - \eta\nabla E_{aug}\big(W(t)\big)$

$\qquad\qquad = \mathrm{W(t)} - \eta(\nabla E_{in}\big(W(t)\big) + \frac{2\lambda}{N}\mathrm{W})$

$\qquad\qquad = \mathrm{W(t)} - \eta * \frac{2\lambda}{N}\mathrm{W(t)} - \eta\nabla E_{in}\big(W(t)\big)$

$\qquad\qquad = \left(1 - \eta * \frac{2\lambda}{N}\right)\mathrm{W(t)} - \eta\nabla E_{in}\big(W(t)\big)$

$$\therefore\ \mathrm{W(t+1)} = \left(1 - \frac{2\eta\lambda}{N}\right)\mathrm{W(t)} - \eta\nabla E_{in}\big(W(t)\big)$$

3.

① $D_{train} = \{(-1,0),\ (\rho,1)\}\ D_{val} = \{(1,0)\}$

解方程式: $\begin{cases} -a1 + b1 = 0 \\ \rho * a1 + b1 = 1 \end{cases}$

得到: $\begin{cases} a1 = \frac{1}{1+\rho} \\ b1 = \frac{1}{1+\rho} \end{cases}$

$e_1 = (h_1(1) - 0)^2$

$\quad = (a1 * 1 + b1 - 0)^2$

$\quad = \left(\frac{2}{1+\rho}\right)^2$

② $D_{train} = \{(-1,0),\ (1,0)\}\ D_{val} = \{(\rho,1)\}$

解方程式: $\begin{cases} -a1 + b1 = 0 \\ a1 + b1 = 0 \end{cases}$

得到: $\begin{cases} a1 = 0 \\ b1 = 0 \end{cases}$

$e_2 = (h_1(\rho) - 1)^2$

$\quad = (a1 * \rho + b1 - 1)^2$

$\quad = 1$

③ $D_{train} = \{(\rho,1),\ (1,0)\}\ D_{val} = \{(-1,0)\}$

解方程式: $\begin{cases}\rho * a1 + b1 = 1 \\ a1 + b1 = 0\end{cases}$

得到: $\begin{cases}a1 = \frac{1}{\rho-1} \\ b1 = \frac{1}{1-\rho}\end{cases}$

$$\begin{aligned}e_3 &= (h_1(-1) - 0)^2 \\ &= (a1 * (-1) + b1 - 0)^2 \\ &= \left(\frac{2}{1-\rho}\right)^2\end{aligned}$$

$$\therefore\ E_{loo} = \frac{1}{3}\sum_{n=1}^{3} e_n = \frac{1}{3} * \left(\frac{4}{(1+\rho)^2} + 1 + \frac{4}{(1-\rho)^2}\right)$$

4.

新的 data set $D_{new}$包含了原始 data set $D_{org}$和虛擬 data set $D_{vrt}$，為了達到 regularization 的目的，$D_{vrt}$中的$\tilde{X} = \sqrt{\lambda}\,\mathrm{I}, \tilde{y} = 0$。

設定：

$$\mathrm{I} = \begin{bmatrix}1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1\end{bmatrix} = [e_1 \quad \dots \quad e_K] \qquad e_k = \begin{bmatrix}0 \\ \vdots \\ 1 \\ \vdots \\ 0\end{bmatrix}$$，即第 k 維的數值為 1，其餘維度數值為 0

$$W_t = \begin{bmatrix}W_t(1,1) \\ W_t(2,1) \\ \vdots \\ W_t(d,1)\end{bmatrix}$$ $W_t(k,1)$為向量$W_t$在(k,1)上的數值

$\because \tilde{X} = \sqrt{\lambda}\,\mathrm{I} \quad \tilde{X} = [\tilde{x}_1, \dots \tilde{x}_K]^T$，其中 K=d，d 為$W_t$的維度

$\therefore \tilde{x}_k = \sqrt{\lambda}\, e_k$

$\because$ linear regression 的 SGD：

$$\begin{aligned}W_{t+1} &= W_t + \eta(-\nabla \mathrm{err}(W_t, x_n, y_n)) \\ &= W_t + \eta\left(-\frac{\partial (y_n - {W_t}^T x_n)^2}{\partial W_t}\right) \\ &= W_t + \eta \cdot 2(y_n - {W_t}^T x_n)(x_n)\end{aligned}$$

Pseudo code:

For i in total iterations:

For $(x_n, y_n)$ in $D_{new}$:

① 若$(x_n, y_n)$來自於$D_{org}$：

$$W_{t+1} = W_t + \eta \cdot 2(y_n - {W_t}^T x_n)(x_n)$$

① 若$(x_n, y_n)$來自於$D_{vrt}$，則$(x_n, y_n) \in \{(\tilde{x}_1, \tilde{y}_1), \dots (\tilde{x}_K, \tilde{y}_K)\}$：

$$\begin{aligned}W_{t+1} &= W_t + \eta \cdot 2(y_n - {W_t}^T x_n)(x_n) \\ &= W_t + \eta \cdot 2(\tilde{y}_k - {W_t}^T \tilde{x}_k)(\tilde{x}_k) \\ &= W_t + \eta \cdot 2(0 - {W_t}^T \sqrt{\lambda}\, e_k)(\sqrt{\lambda}\, e_k) \\ &= W_t - \eta \cdot 2\lambda W_t(k,1) \cdot e_k\end{aligned}$$

上式可理解為：

$$W_{t+1} = \begin{bmatrix} W_t(1,1) \\ \vdots \\ (1-2\eta\lambda)W_t(k,1) \\ \vdots \\ W_t(d,1) \end{bmatrix}$$

即：對$W_t$在$(k,1)$上的數值做縮放操作，縮放的倍數為$(1-2\eta\lambda)$，$W_t$其他維度的數值保持不變。

$$W_t = W_{t+1}$$

由上述流程可知：加入$\tilde{X} = \sqrt{\lambda}\,\mathrm{I}\ \ \tilde{y} = 0$的 virtual data 後，當 scan 到來自$D_{org}$的 data point，演算法會進行一般的 linear regression 的 SGD；當 scan 到來自$D_{vrt}$的 data point，由於 virtual data 數值的特殊性，演算法會對$W_t$中某個維度的數值做縮放的操作。經過一次 iteration 之後，$W_t$中每個維度的數值都經歷過一次縮放操作，所以有達到 regularization 的作用。在 Coursera Q3 中，update rule 的數學式中存在$(1-\frac{2\eta\lambda}{N})W_t$這一項，說明演算法每一次 update 中都會對$W_t$做數值縮放的操作，因此上述兩種方法都有起到 regularization 的作用。

5.

$$\because E_{in}(w) = (sin(ax) - wx)^2 = sin^2(ax) + (wx)^2 - 2wx \cdot sin(ax)$$

$$\therefore \int E_{in}(w)\ dx = \int sin^2(ax) + (wx)^2 - 2wx \cdot sin(ax)\ dx$$

$$= \int \frac{1-cos(2ax)}{2} + (wx)^2 - 2wx \cdot sin(ax)\ dx$$

$$\therefore \int E_{in}(w)\ dx = \left[\frac{1}{2}x - \frac{1}{4a}sin(2ax) + \frac{1}{3}w^2x^3 - 2w(\frac{sin(ax)}{a^2} - \frac{x \cdot cos(ax)}{a})\right]\Bigg|_0^{2\pi}$$

$$= \frac{8}{3}\pi^3 w^2 - 2\left(\frac{sin(2a\pi)}{a^2} - \frac{2\pi \cdot cos(2a\pi)}{a}\right)w - \frac{1}{4a}sin(4a\pi) + \pi$$

$$\int E_{in}(w)\ dx = Aw^2 + Bw + \mathrm{C}$$

$$\because \frac{\partial \int E_{in}(w)\ dx}{\partial w} = 2\mathrm{Aw} + \mathrm{B}$$

$$\because \mathrm{w}_{min} = \underset{w}{\mathrm{argmin}} \int E_{in}(w)\ dx$$

$$\therefore 2\mathrm{Aw}_{min} + \mathrm{B} = 0$$

$$\therefore \mathrm{w}_{min} = \frac{-B}{2A} = \frac{2\left(\frac{sin(2a\pi)}{a^2} - \frac{2\pi \cdot cos(2a\pi)}{a}\right)}{2 \cdot \frac{8}{3}\pi^3}$$

$$= \frac{3 \cdot sin(2a\pi) - 6\pi a \cdot cos(2a\pi)}{8a^2\pi^3}$$

$\therefore$ deterministic noise:

$$|sin(a\pi) - \mathrm{w}_{min}x| = \left|sin(a\pi) - \frac{3 \cdot sin(2a\pi) - 6\pi a \cdot cos(2a\pi)}{8a^2\pi^3}x\right|$$