(2)

(1)

1. Gradient descent for Kernel regression

~~Consid~~ given $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^m$

and $f(x;\theta) = \frac{1}{\sqrt{m}} \theta^T \phi(x)$.

from these two, we can infer that $x \in \mathbb{R}^d$, $\phi(x) \in \mathbb{R}^M$, $\theta \in \mathbb{R}^M$

$$L(\theta) = \frac{1}{2} \| f(X;\theta) - y \|^2$$

where $f(X;\theta) =$ is a vector of predictions $= \begin{bmatrix} f(x_1;\theta) \\ f(x_2;\theta) \\ \vdots \\ f(x_n;\theta) \end{bmatrix}$

(Assume n data points) $(n \times 1)$

$$\frac{\partial f(x;\theta)}{\partial \theta} = \frac{1}{\sqrt{m}} \phi(x) \in \mathbb{R}^M \quad \text{--- (1)}$$

$$\frac{d\theta_t}{dt} = -\frac{\partial L(\theta_t)}{\partial \theta_t} = -\underbrace{\frac{\partial f(X;\theta_t)}{\partial \theta}}_{\substack{\text{Jacobian matrix} \\ \text{of shape } (m \times n)}} \cdot \underbrace{(f(X;\theta_t) - y)}_{\substack{\text{Vector of shape} \\ (n \times 1)}} \quad \text{--- (2)}$$

need to derive. $\frac{d}{dt} f(x;\theta_t)$.

Consider, $\frac{d}{dt} f(x;\theta_t)$

Using chain rule,

$$= \sum_{i=1}^{M} \underbrace{\frac{\partial f(x;\theta_t)}{\partial \theta_i}}_{\text{Scaler}} \cdot \underbrace{\frac{d\theta_i}{dt}}_{\text{Scaler}},$$

$$= \left\langle \underbrace{\frac{\partial f(x;\theta_t)}{\partial \theta}}_{\text{Vector}}, \underbrace{\frac{d\theta}{dt}}_{\text{Vector}} \right\rangle$$

$$= \underbrace{\left(\frac{\partial f(x;\theta_t)}{\partial \theta}\right)^T}_{\text{this is replaced by (1)}} \underbrace{\left(\frac{d\theta}{dt}\right)}_{\text{replace by (2)}}$$

$$= \frac{1}{\sqrt{m}} \phi(x)^T \cdot \underbrace{\left(-\frac{\partial f(X;\theta_t)}{\partial \theta}\right)}_{m\times n \text{ matrix}} \left(f(X;\theta_t) - y\right) \quad \text{— (3)}$$

need to ~~Simp~~ Compute $\frac{\partial f(X;\theta_t)}{\partial \theta}$.

Entries in $\frac{\partial f(X;\theta_t)}{\partial \theta} = \begin{bmatrix} \frac{\partial f(x_1;\theta_t)}{\partial \theta_1} & \frac{\partial f(x_2;\theta_t)}{\partial \theta_2} & \cdots & \frac{\partial f(x_n;\theta_t)}{\partial \theta_1} \\ \vdots & & & \\ \frac{\partial f(x_1;\theta_t)}{\partial \theta_m} & \cdots\cdots & & \frac{\partial f(x_n;\theta_t)}{\partial \theta_m} \end{bmatrix}$

~~Each column~~

take i^{th} column in the above matrix ;

$$\begin{bmatrix} \frac{\partial f(x_i;\theta_t)}{\partial \theta_1} \\ \frac{\partial f(x_i;\theta_t)}{\partial \theta_2} \\ \vdots \\ \frac{\partial f(x_i;\theta_t)}{\partial \theta_m} \end{bmatrix} \rightsquigarrow \text{this can be replaced by, } \frac{1}{\sqrt{m}}\phi(x_i) \quad (\text{from } (1))$$

$$\Rightarrow \frac{\partial f(X;\theta_t)}{\partial \theta} = \underbrace{\begin{bmatrix} | & & | & & | \\ \phi(x_1) & \cdots & \phi(x_i) & \cdots & \phi(x_n) \\ | & & | & & | \end{bmatrix}}_{(m\times n)} \cdot \frac{1}{\sqrt{m}}$$

~~define~~ the above matrix is same as $\left(\frac{1}{\sqrt{m}}\phi(X)\right)$

Substituting in (3)

$$\Rightarrow \boxed{\begin{aligned} \frac{d}{dt} f(x;\theta_t) &= \frac{1}{\sqrt{m}} \cdot \phi(x)^T \frac{-1}{\sqrt{m}} \phi(X) \cdot (f(X;\theta_t) - y) \\ &= -\frac{1}{m} \phi(x)^T \phi(X) \cdot (f(X;\theta_t) - y) \end{aligned}}$$

B.

Need to compute $\frac{d}{dt} f(X;\theta)$.

Consider, $\underbrace{\frac{d}{dt} f(X;\theta)}_{n \times 1 \text{ vector}} = \frac{d}{dt}\begin{bmatrix} f(x_1;\theta) \\ f(x_2;\theta) \\ \vdots \\ f(x_n;\theta) \end{bmatrix} \triangleq \begin{bmatrix} \frac{d}{dt} f(x_1;\theta) \\ \frac{d}{dt} f(x_2;\theta) \\ \vdots \\ \frac{d}{dt} f(x_n;\theta) \end{bmatrix}$ (5)

we ~~can~~ use the result from part A, which is,

$$\frac{d}{dt} f(x;\theta_t) = -\frac{1}{m} \cdot \phi(x)^T \phi(X) \left(f(X;\theta_t) - y \right)$$

here x could be any data point i. So we substitute in (5)

$$\frac{d}{dt} f(X;\theta) = \begin{bmatrix} -\frac{1}{m} \phi(x_1)^T \phi(X) \left(f(X;\theta) - y\right) \\ -\frac{1}{m} \phi(x_2)^T \phi(X) \cdot \left(f(X;\theta) - y\right) \\ \vdots \\ -\frac{1}{m} \phi(x_m)^T \phi(X) \left(f(X;\theta) - y\right) \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{1}{m}\phi(x_1)^T\phi(X) \\ -\frac{1}{m}\phi(x_2)^T\phi(X) \\ \vdots \\ -\frac{1}{m}\phi(x_n)^T\phi(X) \end{bmatrix} \begin{bmatrix} | \\ f(X;\theta)-y \\ | \end{bmatrix}$$

take i th row of this matrix,

$$-\frac{1}{m}\phi(x_i)^T\phi(X)$$

$$= \frac{1}{m}\phi(x_i)^T \begin{bmatrix} | & & | & & | \\ \phi(x_1) & \cdots & \phi(x_j) & \cdots & \phi(x_n) \\ | & & | & & | \end{bmatrix}$$

$$= -\frac{1}{m}\left[\phi(x_i)^T\phi(x_1),\ \phi(x_i)^T\phi(x_2) \cdots \phi(x_i)^T\phi(x_j) \cdots \phi(x_i)^T\phi(x_n)\right]$$

(6)

$$\Rightarrow \frac{d}{dt} f(x;\theta) = \frac{-1}{m} \begin{bmatrix} \phi(x_1)^T \phi(x_1) & \phi(x_1)^T \phi(x_2) & \cdots & \phi(x_1)^T \phi(x_n) \\ \vdots & & & \vdots \\ \phi(x_n)^T \phi(x_1) & \cdots & & \phi(x_n)^T \phi(x_n) \end{bmatrix} \begin{bmatrix} | \\ f(x;\theta)-y \\ | \end{bmatrix}$$

A, $(i,j)^{th}$ element in this matrix,

$$= \frac{1}{m} \phi(x_i)^T \phi(x_j) = k_{i,j}$$

$$\Rightarrow \boxed{\frac{d}{dt} f(x;\theta) = -k\,(f(x;\theta)-y)} \quad \text{— (6)}$$

c) $\frac{d}{dt} v_t = -k v_t$ $\Big\}$ $v_t = e^{-tk} v_0$

to use the above formula, Vector v_t should be same on left & right, demote. $\left(f(x;\theta) - y\right) = \underline{a}$

$$\Rightarrow \frac{d}{dt} f(x;\theta) = \frac{da}{dt}$$

the above,

Substitue in equation 6

$$\Rightarrow \frac{d\underline{a}}{dt} = -k\,\underline{a}$$

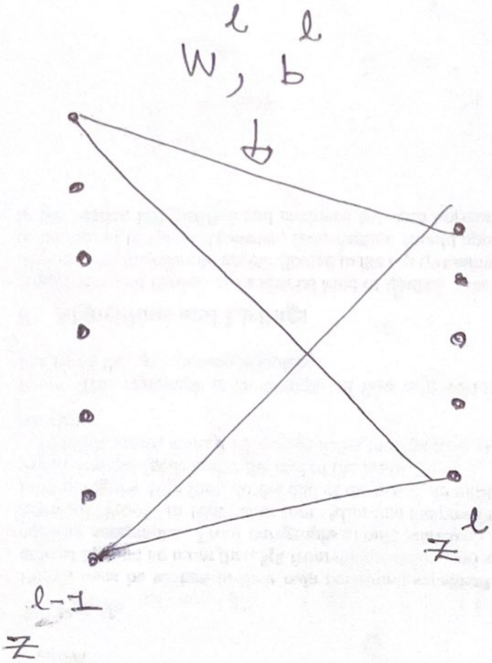
$$\Rightarrow \underline{a} = e^{-tk} a_0$$

$$\Rightarrow \left(f(X;\theta) - y\right) = e^{-tk} a_0$$

at $t=0$ $\quad a = f(X;\theta_0) - y$

$$\Rightarrow f(x;\theta) - y = e^{-tk} \cdot \left(f(X;\theta_0) - y\right)$$

$$\boxed{\Rightarrow f(x;\theta) = y + e^{-tk} \left(f(X;\theta_0) - y\right)}$$



$$z^l(x) = \frac{\sigma_w}{\sqrt{n^{l-1}}} \cdot W^l \sigma\left(z^{l-1}(x)\right) + \sigma_b b^l$$

$$\begin{bmatrix} z_i^{l-1}(x) \\ z_i^{l-1}(x') \end{bmatrix} \sim N\left(0, \begin{bmatrix} \Sigma^{l-1}(x, x) & \Sigma^{l-1}(x', x) \\ \Sigma^{l-1}(x, x') & \Sigma^{l-1}(x', x') \end{bmatrix}\right)$$

given z_i^{l-1} and z_j^{l-1} are independent for $i \neq j$

Consider j^{th} component of $z^l(x)$.

$$z_j^l(x) = \frac{\sigma_w}{\sqrt{n^{l-1}}} \cdot \underbrace{W_j^l}_{\text{first row of } W^l} \sigma\left(z^{l-1}(x)\right) + \sigma_b b_j^l$$

$$Z^{l}_{j}(x) = \frac{\sigma_w}{\sqrt{n^{l-1}}} \cdot \sum_{i=1}^{n^{l-1}} w^{l}_{ij}\; \rho\left(Z^{l-1}_{i}(x)\right) + \sigma_b b^{l}_{j}$$

milarly.

$$Z^{l}_{j}(x') = \frac{\sigma_w}{\sqrt{n^{l-1}}} \cdot \sum_{i=1}^{n^{l-1}} w^{l}_{ij}\; \rho\left(Z^{l-1}_{i}(x')\right) + \sigma_b b^{l}_{j}$$

Qustion is about $E\left(Z^{l}_{j}(x)\right)\left(Z^{l}_{j}(x')\right)$

So consider $\left(Z^{l}_{j}(x)\right)\left(Z^{l}_{j}(x')\right)$

$$= \left(\frac{\sigma_w}{\sqrt{n^{l-1}}} \cdot \sum w^{l}_{ij}\; \rho\left(Z^{l-1}_{i}(x)\right) + \sigma_b b^{l}_{j}\right)$$

$$\left(\frac{\sigma_w}{\sqrt{n^{l-1}}} \cdot \sum w^{l}_{ij}\; \rho\left(Z^{l-1}_{i}(x')\right) + \sigma_b b^{l}_{j}\right)$$

$$= \sigma_b^2 b_j^2 + \frac{\sigma_w^2}{n^{l-1}} \left(\sum_{i=1}^{n^{l-1}} w^{l}_{ij}\; \rho\left(Z^{l-1}_{i}(x)\right)\right)\left(\sum_{i=1}^{n^{l-1}} w^{l}_{ij}\; \rho\left(Z^{l-1}_{i}(x')\right)\right)$$

$$+ \sigma_b b^{l}_{j} \left(\frac{\sigma_w}{\sqrt{n^{l-1}}}\right)\left(\sum_{i=1}^{n^{l-1}} w^{l}_{ij}\; \rho\left(Z^{l-1}_{i}(x)\right) + \sum_{i=1}^{n^{l-1}} w^{l}_{ij}\; \rho\left(Z^{l-1}_{i}(x')\right)\right)$$

$$\sigma_b^2 b_j^{l^2} + \frac{\sigma_w^2}{n^{l-1}}\left[w_{1j}^{l} \, g(z_1^{l-1}(x)) \cdot w_{1j}^{l} \, g(z_1^{l-1}(x')) + w_{1j}^{l} \, g(\cdots) w_{2j}^{l} \, g(\cdot) + \cdots \right] +$$

$$\sigma_b b_j^l \left(\frac{\sigma_w}{\sqrt{n^{l-1}}}\right)\left(\sum_{i=1}^{n^{l-1}} w_{ij}^l \, g(z_i^{l-1}(x)) + \sum_{i=1}^{n^{l-1}} w_{ij}^l \, g(z_i^{l-1}(x')) \right)$$

$$E\left[z_j^l(x) \cdot z_j^l(x') \right] = E\left[\sigma_b^2 b_j^{l^2} \right] + E\left[\quad \right]$$

$$+ E\left[\quad \right]$$

<u>Given</u> $b_j^l \sim N(0,1) \Rightarrow E[b_j^l] = 0, \; E[b_j^{l^2}] = 1$

$w_{ij}^l \sim N(0,1) \Rightarrow E[w_{ij}^l] = 0 \quad E[w_{ij}^{l^2}] = 1$

$$b_j^{l^2} + \left(\frac{\sigma_w^2}{n^{l-1}} \left[w_{1j}^{l} \, g(z_1^{l-1}(x)) \cdot w_{1j}^{l} \, g(z_1^{l-1}(x')) + w_{1j}^{l} \, g(\cdots) \, w_{2j}^{l} \, g(\cdot) + \cdots \right] + \right.$$

$$\sigma_b \, b_j^{l} \left(\frac{\sigma_w}{\sqrt{n^{l-1}}} \right) \left(\sum_{i=1}^{n^{l-1}} w_{ij}^{l} \, g(z_i^{l-1}(x)) + \sum_{i=1}^{n^{l-1}} w_{ij}^{l} \, g(z_i^{l-1}(x'')) \right)$$

$$E\left[z_j^l(x) \cdot z_j^l(x') \right] = E\left[\sigma_b^2 \, b_j^{l^2} \right] + E\left[\quad \right]$$

$$+ E\left[\quad \right]$$

<u>Given</u> $b_j^l \sim N(0,1) \Rightarrow E[b_j^l] = 0, \; E[b_j^{l^2}] = 1$

$w_{ij}^l \sim N(0,1) \Rightarrow E[w_{ij}^l] = 0 \quad E[w_{ij}^{l^2}] = 1$

$$\sigma_b^2 b_j^{l^2} + \left(\frac{\sigma_w^2}{n^{l-1}} \left[w_{1j}^{l}\, \rho(z_1^{l-1}(x)) \cdot w_{1j}^{l}\, \rho(z_1^{l-1}(x')) + w_{1j}^{l}\, \rho(\cdots)\, w_{2j}^{l}\, \rho(\cdot) + \cdots\cdots \right] \right) +$$

$$\sigma_b b_j^l \left(\frac{\sigma^w}{\sqrt{n^{l-1}}} \right) \left(\sum_{i=1}^{n^{l-1}} w_{ij}^l\, \rho(z_i^{l-1}(x)) + \sum_{i=1}^{n^{l-1}} w_{ij}^l\, \rho(z_i^{l-1}(x'')) \right)$$

$$E\left[z_j^l(x) \cdot z_j^l(x') \right] = E\left[\sigma_b^2 b_j^{l^2} \right] + E\left[\quad \right]$$

$$+ E\left[\quad \right]$$

<u>Given</u> $b_j^l \sim N(0,1) \Rightarrow E[b_j^l] = 0, \quad E[b_j^{l^2}] = 1$

$w_{ij}^l \sim N(0,1) \Rightarrow E[w_{ij}^l] = 0 \quad E[w_{ij}^{l^2}] = 1$

~~In the~~

and w_{ij}^{l} and z_i^{l-1} are independent

$$\Rightarrow E\left[w_{ij}^{l}\, z_i^{l-1}\right] = 0$$

or stronger statement $E\left[f(w_{ij}^{l}) \cdot g(z_i^{l-1})\right] = 0$

first term $= E\left[\sigma_b^2\, b_j^{l^2}\right] = \sigma_b^2\, E\left[b_j^{l^2}\right] = \sigma_b^2$

Second term $= E\left[\frac{\sigma_w^2}{n^{l-1}}\left[w_{1j}^{l}\, \varsigma(z_1^{l-1}(x)) \cdot w_{1j}^{l}\, \varsigma(z_1^{l-1}(x')) + w_{1j}^{l}\, \varsigma(z_1^{l-1}(x)) \cdot w_{2j}^{l}\, \varsigma(z_2^{l-1}(x')) + \cdots\right]\right]$

$$= \frac{\sigma_w^2}{n^{l-1}}\left[E\left[\sum_{i=1}^{n^{l-1}} w_{ij}^{l}\, \varsigma(z_i^{l-1}(x)) \cdot \varsigma(z_i^{l-1}(x'))\right] + \left\{ w_{1j}^{l}\, \varsigma(z_1^{l-1}(x))\, w_{2j}^{l}\, \varsigma(z_2^{l-1}(x')) + \cdots \right\}\right]$$

all these are cross terms

$$= \frac{\sigma_w^2}{n^{l-1}} \cdot n^{l-1} \sum_{i=1}^{n} E\left[g(z_i^{l-1}(x)) \cdot g(z_i^{l-1}(x')) \right]$$

$$= \sigma_w^2 \cdot E\left[g(z_i^{l-1}(x))^T g(z_i^{l-1}(x')) \right]$$

third term :- Expectation is zero because of independce.

$$E\left[z_j^l(x)\, z_j^l(x') \right] = \text{first term} + \text{second term} + \text{third}$$

$$= \sigma_B^2 + \sigma_w^2 \cdot E\left[g(z_i^{l-1}(x))^T g(z_i^{l-1}(x')) \right]$$