1. Linear Algebra

(i) $\det A = \underset{m\times m}{U}\ \underset{m\times n}{\Sigma}\ \underset{n\times n}{V^*}$ $\quad V^*$ = conjugate transpose

Define $K_{m\times m}$ → anti diagonal matrix with all entries 1

Note: $B = A^T K$ will give A rotated by 90° as shown.

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \qquad \underset{A^T\ m\times m}{\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ \vdots & \vdots & & \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}} \underset{K\ m\times m}{\begin{bmatrix} 0 & & & 1 \\ & & 1 & \\ & 1 & & \\ 1 & & & 0 \end{bmatrix}} = \underset{B}{\begin{bmatrix} a_{m1} & \cdots & a_{21} & a_{11} \\ \vdots & & \vdots & \vdots \\ a_{mn} & \cdots & a_{2n} & a_{1n} \end{bmatrix}}$$

$\therefore$ eigenvalue (B) = eigenvalues $\left((U\Sigma V^*)^T K \right)$

= eigenvalues $\left(V^{*T} \Sigma^T U^T K \right)$

- Now $\Sigma^T = \Sigma$ since it's rectangular diagonal matrix (with some abuse of notation.)
- Let $U^T K = U'$

= eigenvalue $\left(V^{*T} \Sigma\ U' \right)$

= diagonal elements of Σ

= eigen values of A.

Hence, both A & B have same eigen values.

2 Norms x (n vector), $A_{m\times n}$ ← Given

(a) $\|x\|_\infty \le \|x\|_2$

Let $x = (a_1, a_2 \text{ ——— }, a_m)$

$\|x\|_\infty = \max_k |a_k| \quad k \in \{1, 2, \text{—}, m\} = |a_i|$ (let)

Now

$$a_i^2 \le a_i^2$$

$$\le a_1^2 + a_2^2 + \text{ ——— } + a_i^2 + \text{ ——— } + a_m^2$$

Take squareroot both sides

$$|a_i| \le \sqrt{a_1^2 + a_2^2 + \text{ ——— } + a_m^2}$$

$$\|x\|_\infty \le \|x\|_2 \qquad \text{Hence, Proved}$$

Note equality holds iff $a_k = 0$ $\forall k \neq i$ i.e all vectors x with only 1 non-zero element Eg $x = e_i$

(b) $\|x\|_2 \le \sqrt{n}\, \|x\|_\infty$

Again let $\|x\|_\infty = a_i$ for some $i \in \{1, \text{—}, m\}$

where $x = (a_1, a_2 \text{ ——— }, a_m)$

Now since a_i has highest magnitude

$$a_1^2 \le a_i^2 \quad -①$$
$$a_2^2 \le a_i^2 \quad -②$$
$$\vdots$$
$$a_m^2 \le a_i^2 \quad \text{———} ⓜ$$

Adding ①, ② —— ⓜ

$$a_1^2 + a_2^2 + \text{ ——— } + a_m^2 \le m\, a_i^2$$

$$\Rightarrow \sqrt{a_1^2 + a_2^2 + \text{ ——— } + a_m^2} \le \sqrt{m}\, |a_i|$$

$$\Rightarrow \|x\|_2 \le \sqrt{n}\, \|x\|_\infty$$

Equality $\|x\|_2 = \sqrt{n}\,\|x\|_\infty$

holds iff $a_1 = a_2 = \cdots = a_n$ i.e all elements of x are equal.

Otherwise $\|x\|_2 < \sqrt{n}\,\|x\|_\infty$

Eg $x = (\lambda, \lambda, \ldots, \lambda)$ $\lambda \in \mathbb{R}$

(c) $\|A\|_\infty$ is defined as $\max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty}$ $\quad A_{m\times n}\ x_{n\times 1}\ (Ax)_{m\times 1}$

① – Using part a) $\|Ax\|_\infty \leq \|Ax\|_2$

② – Using part b) $\|x\|_\infty \geq \frac{1}{\sqrt{n}}\|x\|_2$

Combining ① & ② $\Rightarrow \frac{1}{\|x\|_\infty} \leq \frac{\sqrt{n}}{\|x\|_2}$

$$\|A\|_\infty = \max_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} \leq \max_{x\neq 0} \sqrt{n}\,\frac{\|Ax\|_2}{\|x\|_2} \leq \sqrt{n}\,\|A\|_2$$

Hence, Proved.

⊛ Equality is achieved if $\|Ax\|_\infty = \|Ax\|_2$ i.e Ax has exactly one non-zero element. Also $\|x\|_\infty = \frac{\|x\|_2}{\sqrt{2}} \Rightarrow$ All elements of x are exactly same. Hence A has exactly one non-zero row than above conditions are satisfied. (This is one such A)

Eg $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \lambda \\ \lambda \\ \vdots \\ \lambda \end{bmatrix} = \begin{bmatrix} \lambda(a_{11}+a_{12}+\cdots+a_{1n}) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

But we can't put restriction on x. In general if we take with 1 row containing all elements we can show that $\|A\|_\infty = \|Ax\|_2$ if rest of the rows are zero.

Let $A = \begin{bmatrix} \lambda & \lambda & \lambda & \cdots & \lambda \\ \vdots & & 0 & & \vdots \\ \vdots & & 0 & & \vdots \\ \vdots & & 0 & & \vdots \end{bmatrix}_{m \times n}$ (Let $\lambda > 0$)

$\|A\|_\infty = \max$ sum $= n\lambda$

$\|A\|_2 = \sqrt{\text{highest eigen value } (A^T A)}$

$A^T A = \begin{bmatrix} \lambda^2 & \lambda^2 & \cdots & \lambda^2 \\ \lambda^2 & \lambda^2 & \cdots & \lambda^2 \\ \lambda^2 & \lambda^2 & \cdots & \lambda^2 \end{bmatrix}_{n \times n}$ All value $= \lambda^2$

Note Since rank $= 1$ all eigenvalues are zero except 1 i.e corresponding to eigen vector $(1, 1, \cdots 1)^T$

$$\begin{bmatrix} \lambda^2 & \lambda^2 & \cdots & \lambda^2 \\ \lambda^2 & \lambda^2 & \cdots & \lambda^2 \\ \lambda^2 & \lambda^2 & \cdots & \lambda^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = n\lambda^2 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \boxed{n\lambda^2}$$

$$\|A\|_2 = \sqrt{n\lambda^2} = \sqrt{n}\,\lambda$$

$\|A\|_\infty = \sqrt{n}\,\|A\|_2$ iff

$n\lambda = \sqrt{n}\,(\sqrt{n}\,\lambda)$

$= n\lambda$

Hence, our matrix A is a general matrix of size $m \times n$ satisfying

$$\boxed{\|A\|_\infty = \sqrt{n}\,\|A\|_2}$$

(d) $\|A\|_2 \leq \sqrt{m}\, \|A\|_\infty$

$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$

Using ① & ②

Using b) $\|Ax\|_2 \leq \sqrt{m}\, \|Ax\|_\infty$ — ①

Using a) $\|x\|_2 \geq \|x\|_\infty$

$\Rightarrow \frac{1}{\|x\|_2} \leq \frac{1}{\|x\|_\infty}$ — ②

$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \leq \max_{x \neq 0} \frac{\sqrt{m}\, \|Ax\|_\infty}{\|x\|_\infty} \leq \sqrt{m}\, \|A\|_\infty$

Hence, Proved

Equality is attained when

- $\sqrt{m}\, \|Ax\|_\infty = \|Ax\|_2$ i.e Ax has all elements equal.
- $\|x\|_\infty = \|x\|_2$ i.e x has only one non zero element

We can choose A s.t it has all elements in column i having some values, and all other columns 0, where i is given by $\|x\|_\infty = a_i$ where $x = (0, 0, 0 \text{ —— } a_i, \text{ —— } 0)$ i.e non zero element of x's index.

$$\begin{bmatrix} 0 & \lambda & 0 \\ 0 & \lambda & 0 \\ 0 & \lambda & 0 \\ 0 & \lambda & 0 \end{bmatrix} \begin{bmatrix} 0 \\ a_i \\ 0 \\ 0 \\ \vdots \end{bmatrix} = \begin{bmatrix} \lambda a_i \\ \lambda a_i \\ \lambda a_i \\ \lambda a_i \end{bmatrix}$$

A x Ax.

But we can't put a condition on x. In general we can show that if we take $A = \begin{bmatrix} \lambda & 0 & \text{——} & 0 \\ \lambda & 0 & & 0 \\ \lambda & 0 & & 0 \end{bmatrix}$ i.e only first column non zero with equal values, then $\|A\|_2 = \sqrt{m}\, \|A\|_\infty$

1st column non zero with equal values

Let $A = \begin{bmatrix} \lambda & | & | & \cdots & | \\ \lambda & | & | & & | \\ \lambda & | & | & & | \\ | & | & | & & | \\ \lambda & | & | & & | \end{bmatrix}_{m \times n}$ $\lambda > 0$ (let)

$\|A\|_\infty = \lambda$ (max row sum)

$\|A\|_2 = \sqrt{\text{max eigen value}(A^TA)}$

$A^TA = \begin{bmatrix} m\lambda^2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & | & & & | \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$

All elements zero except $(1,1)$.

Clearly all eigenvectors are zero except $(1, 0, 0 \cdots 0)^T$ and eigen value $= m\lambda^2$

$\therefore \|A\|_2 = \sqrt{m\lambda^2} = \sqrt{m}\lambda = \sqrt{m}\,\|A\|_\infty$

Hence A is an eg of general $m \times n$ matrix s.t

$$\boxed{\|A\|_2 = \sqrt{m}\,\|A\|_\infty}$$

Q2

Subgradients

(a) $f(x) = \max_{i=1,2,\ldots,m} (a_i^T x + b_i)$

Find $k \in 1,2,\ldots,m$ s.t

$$f(x) = a_k^T x + b_k$$

then subgradient at that particular $x = \frac{\partial f(x)}{\partial x} = a_k$

(b) $f(x) = \max_{i=1,2,\ldots,m} |a_i^T x + b_i|$

find $k \in 1,2,\ldots,m$ s.t

$$f(x) = |a_k^T x + b_k|$$

If $a_k^T x + b_k \geq 0$ $f(x) = a_k^T x + b_k$ Subgradient $= \frac{\partial f(x)}{\partial x} = a_k$

If $a_k^T x + b_k < 0$ $f(x) = -a_k^T x - b_k$ Subgradient $= \frac{-\partial f(x)}{\partial x} = -a_k$

(c) $f(x) = \sup_{0 \leq t \leq 1} p(t)$ where

$$p(t) = x_1 + x_2 t + \ldots + x_n t^{n-1}$$

Find a t' such that $f(x) = p(t')$ $0 \leq t' \leq 1$

Since $p(t)$ is a polynomial in t we can plot $p(t)$ (If x is given & fixed) in range $[0,1]$ and detect t at which $p(t)$ is maximum. Or we can find t' by finding roots of $p'(t)$ in $t \in [0,1]$. (Assuming $p(t)$ has one, If not maxima is attained at either $t=0$ or $t=1$)

Then Subgradient is given by

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ | \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 1 \\ t' \\ t'^2 \\ \vdots \\ t'^{n-1} \end{bmatrix}$$

(d) $f(x) = x_{[1]} + \_\_\_\_ + x_{[k]}$ $\quad x_{[i]}$: ith largest element of vector x.

Let $x \in \mathbb{R}^n$

Let at any given $x \in \mathbb{R}^n$ $x_{i_1}, x_{i_2}, \longrightarrow, x_{i_k}$ are top k components of vector x. Define set $S = \{i_1, \longrightarrow, i_k\}$

Then a sub-gradient at point x is given by $g \in \mathbb{R}^n$

S.t
$$g_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases} \qquad \forall i \in \{1, 2, \longrightarrow, n\}$$

(e) $f(x) = \inf_{Ay \le b} \|x - y\|_2^2$

Let for a given x, y^* be the point in feasible region i.e $Ay^* \le b$ at which $f(x)$ attains it infimum. then $f(x) = (x-y^*)^T(x-y^*)$

Subgradient of $f(x) = \nabla_x f(x) = 2x - 2y^* = 2(x-y^*)$

To find such y^* we can solve the Dual Problem. Since Slater's condition is satisfied we have "Strict Duality".

Primal Problem $\min \|x-y\|_2^2$ (Note: x is constant)
Subject to $Ay - b \le 0$

$$L(z, y) = (x-y)^T(x-y) + z^T(Ay-b) = x^Tx - x^Ty - y^Tx + y^Ty + z^TAy - z^Tb$$

$g(z) = \min_y L(z,y)$ To find $g(z)$ let's find. $\nabla_y L(z,y) = -2x + 2y + A^Tz = 0$

$\Rightarrow y = x - \frac{1}{2}A^Tz$ ← substitute it back into L to get $g(z)$

$$g(z) = \left(\frac{1}{2}A^Tz\right)^T x \frac{1}{2}A^Tz + z^T\left(\frac{1}{2}Ax - \frac{1}{2}AA^Tz - b\right) = -\frac{1}{4}z^TAA^Tz + z^TAx - z^Tb$$

Dual Problem is $\max_z g(z)$ subject to $z \ge 0$. Let z^* is soln to dual problem, then optimal value of primal problem i.e $f(x) = \left(-\frac{1}{2}z^{*T}AA^Tz^* - z^{*T}Ax - z^{*T}b\right)$ — (1)

By weak duality we have $f(x') \ge -\frac{1}{2}z^{*T}AA^Tz^* - z^{*T}Ax' - z^{*T}b$ — (2)

$\therefore f(x') - f(x) \ge z^{*T}A(x-x')$ Hence $(z^{*T}A)$ is Subgradient at x.

Now if y^* & z^* are optimal soln then by KKT

$$\nabla_y L(y^*, z^*) = 0$$

$$\Rightarrow \boxed{y^* = x - \frac{1}{2} A^T z^*} \rightarrow \text{Hence we found one such } y.$$

Note: Subgradient = $\boxed{A^T z^* = 2(x - y^*)}$

Q3 Probability

1. $E(I/X=x) = P(I=1/X=x) \times 1 + P(I=0/X=x) \times 0$

$= P(I=1/X=x)$

$= P(Z<x/X=x)$

$= P(Z<x)$ since $Z \sim N(0,1)$

$= \phi(x)$

2. ⊛ $E(I) = P(I=1) \times 1 + P(I=0) \times 0 = P(I=1) = P(Z<X)$ — ①

⊛ Using $E[X] = E[E[X|Y]]$

$E[I] = E(E(I/X))$ Using Part 1.) $E(I/X) = \phi(X)$

$= E(\phi(X))$ — ②

Using ① & ②

$$E[\phi(X)] = P(Z<X)$$

3. $E(\phi(X)) = P(Z<X) = P(X-Z>0)$

Now $X \sim N(\mu, 1)$

$Z \sim N(0,1)$

$-Z \sim N(0,1)$

$\Rightarrow X - Z \sim N(\mu, 2)$

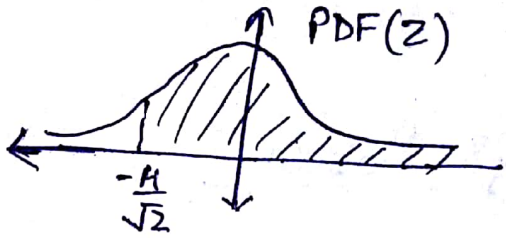
$\Rightarrow \frac{(X-Z)-\mu}{\sqrt{2}} \sim N(0,1)$

$$\therefore E(\phi(x)) = P(X-2>0)$$

$$= P\left(\frac{(X-2)-\mu}{\sqrt{2}} > -\frac{\mu}{\sqrt{2}}\right)$$

↳ It follows standard normal distribution

$$= P\left(Z > -\frac{\mu}{\sqrt{2}}\right)$$



Using Symmetry

$$= P\left(Z < \frac{\mu}{\sqrt{2}}\right)$$

$$= \phi\left(\frac{\mu}{\sqrt{2}}\right)$$

Q4 Machine Learning

4.1 PCA

1. $\max_{u:\|u\|_2=1} \tilde{V}[u^T X]$

Given $\frac{1}{N}\sum_{i=1}^{N} x_i = 0 \quad i = 1, 2, \ldots, N$

- Projected points are $(u_T x_1, u_T x_2, \ldots, u_T x_N)$
- Avg of projected points $\frac{1}{N}\left(u_T \sum_{i=1}^{N} x_i\right) = 0$

$$\left(\therefore \tilde{V}[u^T X] = \sum_{i=1}^{N} \frac{\|u^T x_i - 0\|_2^2}{N}\right) \quad \max_{u:\|u\|_2=1} \sum_{i=1}^{N} \frac{(u^T x_i)(u^T x_i)^T}{N}$$

$$= \max_{u:\|u\|_2=1} u^T \left(\sum_{i=1}^{N} \frac{x_i x_i^T}{N}\right) u$$

$$\therefore \Sigma = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T$$

4.1 PCA

2.

$$\min_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N} \| x_i - uu^T x_i \|_2^2$$

$$= \min_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N} (x_i - uu^T x_i)^T (x_i - uu^T x_i)$$

$$= \min_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N} x_i^T x_i - x_i^T uu^T x_i - x_i^T uu^T x_i + x_i^T u \underbrace{u^T u}_{1} u^T x_i$$

Using Fact that $u^T u = 1$

and $x_i^T u = u^T x_i$ is a scalar

$$= \min_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N} x_i^T x_i - (u^T x_i)^2$$

Ignoring $\sum_{i=1}^{N} x_i^T x_i$ since it doesn't depend on u

$$= \min_{u:\|u\|_2=1} -\frac{1}{N}\sum_{i=1}^{N} (u^T x_i)^2$$

$$= \max_{u:\|u\|_2=1} \frac{1}{N}\sum_{i=1}^{N} (u^T x_i)^2$$ ← Same as Part 1 i.e variance term

$$= \max_{u:\|u\|_2} u^T \Sigma u \quad \text{where} \quad \boxed{\Sigma = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T}$$

4.2 Bias Variance Trade Off

$E_D(J(\theta)) = ?$

Bias $= \bar{h_\theta}(x) - y$ where $\bar{h_\theta}(x) = E_D[h_\theta(x)]$

Variance $= E_D\left((h_\theta(x) - \bar{h_\theta}(x))^2\right)$

$= E_D\left(h_\theta^2(x) - 2\bar{h_\theta}(x)\, h_\theta(x) + \bar{h_\theta}^2(x)\right)$

$= E_D\left(h_\theta^2(x)\right) - \left(\bar{h_\theta}(x)\right)^2$

a) $E_D(J(\theta)) = E_D\left((y - h_\theta(x))^2\right)$

$= E_D\left(y^2 + h_\theta^2(x) - 2y h_\theta(x)\right)$

$= y^2 + E_D\left(h_\theta^2(x)\right) - 2y\bar{h_\theta}(x)$

Add & Subtract $(\bar{h_\theta}(x))^2$

$= \underbrace{\left[y^2 - 2y\bar{h_\theta}(x) + (\bar{h_\theta}(x))^2\right]}_{\text{Bias}^2} + \underbrace{E_D(h_\theta^2(x)) - (\bar{h_\theta}(x))^2}_{\text{Variance}}$

$= \text{Bias}^2 + \text{Variance}$.

b) Bias $= \bar{h_\theta}(x) - f(x)$ Variance $= E_D\left(h_\theta^2(x)\right) - \left(\bar{h_\theta}(x)\right)^2$

$E(J(\theta)) = E_D\left((f(x)+\varepsilon - h_\theta(x))^2\right) = E_D\left((f(x)+\varepsilon)^2 + h_\theta^2(x) - 2h_\theta(x)(f(x)+\varepsilon)\right)$

$= E_D\left(f(x)^2 + \varepsilon^2 + 2\varepsilon f(x)\right) + E_D\left(h_\theta^2(x)\right) - 2E\left(h_\theta(x) f(x)\right) - 2E\left(\varepsilon h_\theta(x)\right)$

Note $f(x)$ is deterministic & not random ε is $N(0, \sigma^2)$

$\therefore E(\varepsilon) = 0$ Variance $= \sigma^2 = E((\varepsilon - 0)^2) = E(\varepsilon^2)$

$= f(x)^2 + E_D(\varepsilon^2) + 2f(x)\underbrace{E(\varepsilon)}_{0} + E_D\left(h_\theta^2(x)\right) - 2f(x)\bar{h_\theta}(x) - 2\underbrace{E(\varepsilon)}_{0}\bar{h_\theta}(x)$

Add & subtract $(\bar{h_\theta}(x))^2$

$= \left[f(x)^2 - 2f(x)\bar{h_\theta}(x) + (\bar{h_\theta}(x))^2\right] + \left(E_D(h_\theta^2(x)) - (\bar{h_\theta}(x))^2\right) + E(\varepsilon^2)$

$= \text{Bias}^2 + \text{Variance} + \sigma^2$

4.3 Kernelizing the Perceptron

$$\theta^{i+1} = \theta^i + \alpha \left[y^{i+1} - h_{\theta^i}(x^{i+1}) \right] x^{i+1}$$

Note: we have find a way to update our parameters as follows

$$\theta^{i+1} = \theta^i + \alpha \left[y^{i+1} - h_{\theta^i}(\phi(x^{i+1})) \right] \phi(x^{i+1})$$

Note: $y^{i+1} \in \{0, 1\}$

$$h_{\theta^i}(\phi(x^{i+1})) = g(\theta^{iT} \phi(x^{i+1})) \in \{0, 1\}$$

$\therefore$ $y^{i+1} - h_{\theta^i}(\phi(x^{i+1})) \in \{-1, 0, 1\}$

Let's call it error

$$\therefore \quad \theta^{i+1} = \begin{cases} \theta^i - \alpha\phi(x^{i+1}) & \text{error} = -1 \\ \theta^i & \text{error} = 0 \\ \theta^i + \alpha\phi(x^{i+1}) & \text{error} = 1 \end{cases}$$

Thus $\theta^{i+1} = \theta^i + \beta_{i+1} \phi(x^{i+1})$ where $\beta_i \in \{-1, 0, 1\}$

$$\theta^i = \theta^{i-1} + \beta_i \phi(x^i)$$

$$\vdots$$

$$\theta^1 = \theta^0 + \beta_1 \phi(x_1)$$

($\theta^0 = \vec{0}$ given)

Adding we get

$$\theta^{i+1} = \beta_1 \phi(x_1) + \beta_2 \phi(x_2) + \cdots + \beta_{i+1} \phi(x^{i+1})$$

where β_i's $\in \{-1, 0, 1\}$

4.3

(1) From derivation above we can clearly see that $\theta^{(i)}$ can be represented as (implicitly) the linear combinations of $\phi(x^1), \phi(x^2) \text{ ——— }, \phi(x^i)$ i.e

$$\theta^{(i)} = \sum_{k=1}^{i} \beta_k \phi(x^k) \quad \text{where } \beta_k \in \{0, 1, -1\}$$

Hence θ can be represented as list of β's. The value $\theta^{(0)}$ is initialized to $\vec{0}$ and can be thought as base case where the summation ends as there are no summation terms included in $\theta^{(0)}$.

(Note: this was helpful end recursion / act as base case for $\theta^{(1)} = \theta^{(0)} + \beta_1 \phi(x^1)$)

→ We can say the list of β's for $\theta^{(0)}$ is empty.

→ $\theta^i \equiv \{\beta_1, \beta_2 \text{ ——— } \beta_i\}$

(2) How to predic $h_{\theta^i}(\phi(x^{i+1}))$

$$= g\left((\theta^i)^T \phi(x^{i+1})\right)$$

Since $\theta^i = \sum_{k=1}^{i} \beta_k \phi(x^k) \quad \left(\theta^{iT} = \sum_{k=1}^{i} \beta_k \phi^T(x^k)\right)$

$$= g\left(\sum_{k=1}^{i} \beta_k \underbrace{\phi^T(x^k)\, \phi(x^{i+1})}\right) = g\left(\sum_{k=1}^{i} \beta_k K(x^k, x^{i+1})\right)$$

This dot product in high dimensional space can be computed as sum function of dot products on x^k, x^{i+1} in orginal space. Let that function be K.

Note: for prediction on any new point we only need βs & K. i.e θ^i & K.

(3) Since we are representing θ as list of β's

i.e

$$\theta^i = \{\beta_1, \beta_2 \text{ ——— } , \beta_i\} \quad \text{where } \beta_k \in \{-1, 0, 1\} \ \forall k$$

Updating θ to θ^{i+1} means getting new value of β i.e β_{i+1}

Now we know that

$$\theta^{i+1} = \theta^i + \alpha \underbrace{\left[y^{i+1} - h_{\theta^i}\left(g\phi(x^{i+1})\right)\right]}_{\text{This gives us } \beta} \phi(x^{i+1})$$

$$\therefore \boxed{\beta^{i+1} = y^{i+1} - h_{\theta^i}\left(g\phi(x^{i+1})\right)}$$

where 2nd term is calculated using method in part 2 using $\{\beta_1, \beta_2 \rightarrow \beta_i\}$ & Kernel function K.

$$\therefore \theta^{i+1} = \{\beta_1, \beta_2, \text{ ——— } \rightarrow \beta_i, \beta_{i+1}\}$$