Homework 3

B99202064 Kuan Hou Chan

1 Ans: (C) 100

$\sigma^2(1-\frac{d+1}{N}) > 0.008$
$\Rightarrow 1-\frac{d+1}{N} > \frac{0.008}{\sigma^2}, hence, N > \frac{d+1}{1-\frac{0.008}{\sigma^2}} = 45$

2 Ans: (A,D,E)

1. $H^2 = H*H = (\chi(\chi^T\chi)^{-1}\chi^T)*(\chi(\chi^T\chi)^{-1}\chi^T) = \chi(\chi^T\chi)^{-1}(\chi^T\chi)(\chi^T\chi)^{-1}\chi^T = H$
$\Rightarrow H^3 = H^2*H = H*H = H$, and for $H^{1126} = H$. Hence (e) is true.
2. Let $HW = \lambda W$, multiply by H $\Rightarrow H^2W = \lambda HW = \lambda(\lambda W)$
$\Rightarrow \because H^2 = H, \therefore H^2W = HW = \lambda W$, hence $\lambda^2 W = \lambda W$
$\therefore \lambda(\lambda-1)W = 0 \Rightarrow \lambda = 0/\lambda = 1$. Hence (c) is false and (a) is true since all eigenvalues of H are non-negative.
3. (d) is true since $tr(H) = tr((\chi(\chi^T\chi)^{-1}\chi^T)) = tr(((\chi^T\chi)^{-1}\chi^T\chi)) = tr(I) = d+1 (\because tr(AB) = tr(BA))$
And trace is the sum of diagonal elements equals the sums of eigenvalues since diagonalize H to
$tr(H) = tr(S\Lambda S^{-1}) = tr(S^{-1}S\Lambda) = tr(\Lambda) = \lambda_1 + \lambda_2 + ...$ Hence tr(h) = d+1 means we have d+1 eigenvalues are 1.
4. $\because \lambda = 0/\lambda = 1$, if H is always invertible then $\lambda \neq 0$ since invertible theorem said it can't exist any trivial solution for invertible matrix. Hence, (b) is false.

3 Ans: (A,B,E)

** The case discuss below also can use limit theorem to compare which is upper bound
(a) By consider extreme case that $W^TX = 0.0001$ and $y = -1$,
then $err(W) = max(0, 1-(-1)*(0.0001)) = 1.0001 > [sign(0.0001) \neq -1] = 1$. Vice versa for $W^TX = -0.0001$ and $y = 1$.
And $err(W) = max(0, 1-(-1)*(0)) = 1 == [sign(0) \neq -1] = 1$
True for $W^TX = 0.0001$ and $y = 1$,
$err(W) = max(0, 1-(1)*(0.0001)) = 0.9999 > [sign(0.0001) = 1] = 0$
Hence, (a) is upper bound.
Since (b) is square of (a), then it is also an upper bound
(c) By consider extreme case that $W^TX = 0.0001$ and $y = -1$,
then $err(W) = max(0, -(-1)*(0.0001)) = 0.0001 < [sign(0.0001) \neq -1] = 1$. Vice versa for $W^TX = -0.0001$ and $y = 1$.
Hence, (c) is not an upper bound.
(d) By consider extreme case that $W^TX = 0.0001$ and $y = -1$,
then $err(W) = \theta(-(-1)*(0.0001)) = 0.5 < [sign(0.0001) \neq -1] = 1$. Vice versa for $W^TX = -0.0001$ and $y = 1$.
Hence, (d) is not an upper bound.
(e) By consider extreme case that $W^TX = 0.0001$ and $y = -1$,
then $err(W) = exp(-(-1)*(0.0001)) = 1.0001 > [sign(0.0001) \neq -1] = 1$. Vice versa for $W^TX = -0.0001$ and $y = 1$.
And $err(W) = exp(-(-1)*(0)) = 1 == [sign(0) \neq -1] = 1$
True for $W^TX = 0.0001$ and $y = 1$,
$err(W) = exp(-(1)*(0.0001)) = 0.9999 > [sign(0.0001) = 1] = 0$
Hence, (e) is upper bound.

4 Ans: (B,D,E)

(a) $\lim_{h\to 0}\frac{max(0,1-y(c+h)x)-max(0,1-ycx)}{h}$
when $ycx > 1$, $err(w) = \lim_{h\to 0}\frac{0-0}{h} = 0$
when $ycx < 1$, $err(w) = \lim_{h\to 0}\frac{(1-ycx-yhx)-(1-ycx)}{h} = \lim_{h\to 0}\frac{-yhx}{h} = -yx$

$$err(w) = \begin{cases} 0 & \text{if } ycx > 1 \\ -yx & \text{if } ycx < 1 \end{cases}$$

Hence err(w) of (a) can't be differentiable at $yx = 1$ since ycx = 1 will be incosistent.
(b) $\lim_{h\to 0}\frac{(max(0,1-y(c+h)x))^2-(max(0,1-ycx))^2}{h}$
when $ycx > 1$, $err(w) = \lim_{h\to 0}\frac{0-0}{h} = 0$
when $ycx < 1$, $err(w) = \lim_{h\to 0}\frac{(1-ycx-yhx)^2-(1-ycx)^2}{h} = \lim_{h\to 0}\frac{-2yhx+2y^2x^2ch+y^2x^2h^2}{h} = 2yx(ycx-1)$

$$err(w) = \begin{cases} 0 & \text{if } ycx > 1 \\ 2yx(ycx-1) & \text{if } ycx < 1 \end{cases}$$

Hence err(w) of (b) can be differentiable everywhere.
(c) Same as (a).
(d) and (e) can be differentiable everywhere because of exponential function can be differentiable everywhere.

5 Ans: (C) $err(w) = max(0, -yW^Tx)$

Since PLA is adjust error by its vector of weight when detect error,
then option (a) and (b) will limit the error within $[0, 1]$
Means that although it found the error but it only can update weight within $[0, 1]$
On the other hand option (d) and (e) will always update weight nontheless it is error or not since it don't have any mechanism to distinguish error.
Hence option (c) fulfill the condition that it will update weight by the length of vector when error happen, then (c) will result in PLA.

6 Ans: (D) $(-2, 0)$

For the following question of 6-10:

$$E(u,v) = e^u + e^{2v} + e^{uv} + u^2 - 2uv + 2v^2 - 3u - 2v$$
$$\frac{\partial E}{\partial u} = e^u + ve^{uv} + 2u - 2v - 3$$
$$\frac{\partial E}{\partial v} = 2e^{2v} + ue^{uv} - 2u + 4v - 2$$
$$\frac{\partial^2 E}{\partial u^2} = e^u + v^2e^{uv} + 2$$
$$\frac{\partial^2 E}{\partial v^2} = 4e^{2v} + u^2e^{uv} + 4$$
$$\frac{\partial^2 E}{\partial v\partial u} = \frac{\partial^2 E}{\partial u\partial v} = vue^{uv} + e^{uv} - 2$$

$\frac{\partial E}{\partial u}|_{u=0,v=0} = 1 + 0 + 0 - 0 - 3 = -2$
$\frac{\partial E}{\partial v}|_{u=0,v=0} = 2 + 0 - 0 + 0 - 2 = 0$

7 Ans: (C) 2.825

$$u_{t+1}, v_{t+1} = (u_t, v_t) - \eta \nabla E(u_t, v_t)$$

$(u_1, v_1) = (0, 0) - 0.01 * (-2, 0) = (0.02, 0)$
$(u_2, v_2) = (0.02, 0) - 0.01 * (-1.93, -0.02) = (0.0393, 0.0002)$
$(u_3, v_3) = (0.0393, 0.0002) - 0.01 * (-1.88, -0.038) = (0.0568, 0.00058)$
$(u_4, v_4) = (0.0568, 0.00058) - 0.01 * (-1.828, -0.052) = (0.07508, 0.0011)$
$(u_5, v_5) = (0.07508, 0.0011) - 0.01 * (-1.773, -0.0663) = (0.09281, 0.00176)$
$\therefore E(u_5, v_5) = 2.844$

8 Ans: (B) $(1.5, 4, -1, -2, 0, 3)$

$$\widehat{E}_2(\triangle u, \triangle v) = b_{uu}(\triangle u)^2 + b_{vv}(\triangle v)^2 + b_{uv}(\triangle u)(\triangle v) + b_u(\triangle u) + b_v(\triangle v) + b$$

Using Taylor expansion to $E(u+ \triangle u, v+ \triangle v)$, we can obtain:

$$f(x_0 + \delta x) \approx f(x_0) + \nabla f * \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} + \frac{1}{2}(\delta x, \delta y)[\nabla^2] \begin{bmatrix} \delta x \\ \delta y \end{bmatrix}$$

$\therefore \widehat{E}_2(\triangle u, \triangle v) = E(0,0) + (u-0)\frac{\partial E}{\partial u}|_{u=0,v=0} + (v-0)\frac{\partial E}{\partial v}|_{u=0,v=0} +$
$\frac{1}{2!}[(u-0)^2 \frac{\partial^2 E}{\partial u^2}|_{u=0,v=0} + (v-0)^2 \frac{\partial^2 E}{\partial v^2}|_{u=0,v=0} + 2(u-0)(v-0)\frac{\partial^2 E}{\partial v \partial u}|_{u=0,v=0}]$
$\therefore E(0,0) = 3,$
$\frac{\partial E}{\partial u}|_{u=0,v=0} = -2,$
$\frac{\partial E}{\partial v}|_{u=0,v=0} = 0,$
$\frac{\partial^2 E}{\partial u^2}|_{u=0,v=0} = 3,$
$\frac{\partial^2 E}{\partial v^2}|_{u=0,v=0} = 8,$
$\frac{\partial^2 E}{\partial v \partial u}|_{u=0,v=0} = -1$

9 Ans: (A) $-(\nabla^2 E(u,v))^{-1} \nabla E(u,v)$

$$\widehat{E}_2 = E(u,v) + \nabla E(u,v) * (\triangle u, \triangle v) + \frac{1}{2}\nabla^2 E(u,v) * (\triangle u, \triangle v)^2$$

For minimize $\widehat{E}_2$, we differentiate it and let it equal to zero:
$\lim_{(\triangle u \to 0, \triangle v \to 0)} \frac{\widehat{E}_2 - E}{((\triangle u, \triangle v))} = \lim_{(\triangle u \to 0, \triangle v \to 0)} \frac{\nabla E * (\triangle u, \triangle v) + \frac{1}{2}\nabla^2 E * (\triangle u, \triangle v)^2}{(\triangle u, \triangle v)} = \lim_{(\triangle u \to 0, \triangle v \to 0)} \nabla E + \nabla^2 E * (\triangle u, \triangle v) = 0$

$\therefore \nabla^2 E * (\triangle u, \triangle v) = -\nabla E$
Hence,

$$(\triangle u, \triangle v) = -(\nabla^2 E(u,v))^{-1} \nabla E(u,v)$$

10 Ans: (C) 2.361

$$\nabla^2 E(u,v) = \begin{bmatrix} e^u + v^2 e^{(uv)} + 2 & -2 + e^{(uv)} + uve^{(uv)} \\ -2 + e^{(uv)} + uve^{(vu)} & 4e^{(2v)} + u^2 e^{(uv)} + 4) \end{bmatrix}$$

By wolfram alpha:

$$(\nabla^2 E)^{-1} \nabla E = \begin{bmatrix} \frac{((u^2 e^{(uv)} + 4e^{(2v)} + 4)(ve^{(uv)} + 2u + e^u - 2v - 3))}{((u^2 e^{(uv)} + 4e^{(2v)} + 4)(v^2 e^{(uv)} + e^u + 2) - (uve^{(uv)} + e^{(uv)} - 2)^2)} + \frac{((ue^{(uv)} - 2u + 2e^{(2v)} + 4v - 2)(uv(-e^{(uv)}) - e^{(uv)} + 2))}{((u^2 e^{(uv)} + 4e^{(2v)} + 4)(v^2 e^{(uv)} + e^u + 2) - (uve^{(uv)} + e^{(uv)} - 2)^2)} \\ \frac{((ve^{(uv)} + 2u + e^u - 2v - 3)(uv(-e^{(uv)}) - e^{(uv)} + 2))}{((u^2 e^{(uv)} + 4e^{(2v)} + 4)(v^2 e^{(uv)} + e^u + 2) - (uve^{(uv)} + e^{(uv)} - 2)^2)} + \frac{((ue^{(uv)} - 2u + 2e^{(2v)} + 4v - 2)(v^2 e^{(uv)} + e^u + 2))}{((u^2 e^{(uv)} + 4e^{(2v)} + 4)(v^2 e^{(uv)} + e^u + 2) - (uve^{(uv)} + e^{(uv)} - 2)^2)} \end{bmatrix}$$

By using $(u_{t+1}, v_{t+1}) = (u_t, v_t) - (\nabla^2 E)^{-1} \nabla E$, iterate it 5 times from $(0,0)$ and will obtain $E(u_5, v_5) = 2.361$

11 Ans: (E) $X_1, X_2, X_3, X_4, X_5, X_6$

We have 6 points, the quadratic form $d_{vc} = 6$, the linear form $d_{vc} \leq 3$, the constant form $d_{vc} = 1$
By Homework 2, we found that
$max\{d_{vc}(H_k)\}_{k=1}^K \leq d_{vc}(\cap_{k=1}^K H_k) \leq K - 1 + \sum_{k=1}^K d_{vc}(H_k)$,
then we can shattered $6 + 3 + 1 = 10$ inputs. Hence the biggest subset will be $X_1, X_2, X_3, X_4, X_5, X_6$.

12 Ans: (D) ∞

If the transformer peeks the data and include all posible Φ, then will have a huge hypothesis come up with the data set. Then the d_{vc} not only the data set itself create, but also will be charged of the d_{vc} we create in our mind.
Hence, we pay the price and the $d_{vc} = \infty$.

13 Ans: (C) 0.5

14 Ans: (A)

15 Ans: (A)

16 Ans: (D) $\frac{1}{N}\sum_{n=1}^{N}(ln(\sum_{i=1}^{K} exp(W_i^T X_n)) - W_{y_n}^T X_n)$

$\because h_y(x) = \frac{exp(W_y^T X_n)}{\sum_{i=1}^{K} exp(W_i^T X_n)}$,
$max(likelihood(w)) \propto \prod_{n=1}^{N} h_y(X) \Rightarrow max(likelihood(w)) \propto ln \prod_{n=1}^{N} \frac{exp(W_y^T X_n)}{\sum_{i=1}^{K} exp(W_i^T X_n)}$
For minimize $\Rightarrow min_w \frac{1}{N}\sum_{n=1}^{N}[-ln(h_y(x))] = min_w \frac{1}{N}\sum_{n=1}^{N}[ln[\frac{exp(\sum_{i=1}^{K} exp(W_i^T X_n))}{W_y^T X_n}]]$
$\Rightarrow min_w \frac{1}{N}\sum_{n=1}^{N}[ln(\sum_{i=1}^{K} exp(W_i^T X_n)) - W_{y_n}^T X_n]$

17 Ans: (C) $\frac{1}{N}\sum_{n=1}^{N}((h_i(x_n) - [y_n = i])x_n)$

$\frac{\partial E_{in}}{\partial w_i} = \frac{1}{N}\sum_{n=1}^{N}(\frac{\sum_{i=1}^{K} exp(W_i^T X_n) * \frac{\partial}{\partial w_i}(w_i^T X_n)}{\sum_{i=1}^{K} exp(W_i^T X_n)} - x_n \frac{\partial(w_{y_n}^T)}{\partial w_i})$
Hence, $\frac{\partial E_{in}}{\partial w_i} = \frac{1}{N}\sum_{n=1}^{N}((h_i(x_n) - [y_n = i])x_n)$

18 Ans :(A) 0.475

19 Ans :(D) 0.22

20 Ans :(A) 0.473

21 Least number of queries: N+1

$RMSE = 0 = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - h(x_n))^2}$
The dimension of test set will be N + 1 since x_0 counted in.
For RMSE = 0, we must have the term $\sum_{n=0}^{N}(y_n - h(x_n))^2 = 0$,
implied that for each n, $(y_n - h(x_n)) = 0$.
Hence we can query RMSE(h) for some h by (N+1) times to figure out each term of $(y_n - h(x_n))$ to become zero.
And finally we can construct a hypothesis g with RMSE(g)=0.

22 Least number of queries: 1

Let RMSE(h) = Z, expand RMSE equation, we obtain that:
$Z = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - h(x_n))^2} \Rightarrow Nz^2 = (y_1 - h(x_1))^2 + (y_2 - h(x_2))^2 + ...$
$\Rightarrow NZ^2 = y_1^2 - 2y_1 h(x_1) + h(x_1)^2 + y_2^2 - 2y_2 h(x_2) + h(x_2)^2 + ...$
$\Rightarrow NZ^2 = \sum_{n=1}^{N}(y_n^2 + h(x_n)^2) - 2(\overrightarrow{h^T}\overrightarrow{y})$
$\Rightarrow \overrightarrow{h^T}\overrightarrow{y} = \frac{1}{2}(\sum_{n=1}^{N}(y_n^2 + h(x_n)^2)) + NZ^2$
Since we know $h(x_n)$ and can query one times of RMSE to get Z, then we just need to compute $\overrightarrow{h^T}\overrightarrow{y}$ and just query one time only.

23 Least number of queries: K

Let RMSE(H) = Z, take gradient of RMSE equation, we obtain that:
$\frac{\partial}{\partial w_k} Z = 0 = \frac{1}{2} * Z^{-\frac{1}{2}} * \frac{1}{N} * 2 * \sum_{n=1}^{N} (y_n - \sum_k w_k h_k(x_n)) * (\sum_k h_k(x_n))$
square, $\Rightarrow 0 = Z^{-1} * \frac{1}{N^2} * [\sum_{n=1}^{N} (y_n - \sum_k w_k h_k(x_n))]^2 * (\sum_k h_k(x_n))^2$
$\Rightarrow 0 = \frac{1}{Z} * \frac{1}{N^2} * N^2 z^4 * (\sum_k h_k(x_n))^2$
$\Rightarrow 0 = Z^3 * (\sum_k h_k(x_n))^2$
Then we only need to solve the equation $Z^3 = 0$ to minimize the RMSE(H)
By Q22, there's have k of w's that we need to find and it just need query $1 * k = k$ times to solve the w out.