Machine Learning Final Exam Date: Jan 7, 2021 Time: 9:10-12:10

Name: Leung Ko Tsun (梁高浚) SID: 7109230 3

監考：陳岩

Q6 a). $L(\alpha, b, \xi, \omega, \beta, \gamma) = \frac{1}{2}\sum_{i=1}^{N}\alpha_i^2 + \sum_{i=1}^{N} C_i\xi_i + \sum_{i=1}^{N}\omega_i(1-\xi_i - y_i(\sum_{j=1}^{N}\alpha_j y_j x_i \cdot x_j + b))$

$+\sum_{i=1}^{N}\beta_i(-\alpha_i) + \sum_{i=1}^{N}\gamma_i(-\xi_i)$

b). By strong Duality theorem, if X is a nonempty convex set in $\mathbb{R}^N$.

if f, g_1, g_2, g_3 are convex, and $\exists \hat{x} \in X$ s.t. $g_{1,i}(\hat{x}) < 0$, $g_{2,i}(\hat{x}) < 0$ and $g_{3,\bar{i}}(\hat{x}) < 0$, then

$\inf\{f(x): x \in X, g_1(x) \le 0, g_2(x) \le 0, g_3(x) \le 0\} = \sup\{\theta(\omega, \beta, \gamma): \omega \ge 0, \beta \ge 0, \gamma \ge 0\}$, $\omega, \beta, \gamma \in \mathbb{R}^N$

We want to ensure these conditions are satisfied in order to apply strong Duality theorem.

Obviously, X is nonempty. X is also convex set as $tx + (-t)x$ is contained in X.

for function f, $\sum_{i=1}^{N}\alpha_i^2$ is ~~convex~~ quadratic hence convex, $\sum_{i=1}^{N} C_i\xi_i$ is affine hence convex.

for function $g_{1,i}$, it is also convex as second derivative > 0

, $g_{2,i}, g_{3,i}$ they are also convex as they are just affine functions.

and $\exists \hat{x} \in X$ s.t. $g_{1,i}(\hat{x}) < 0$, ~~$g_{1,i}(\hat{x}) < 0$ and $g_{3,i}(\hat{x}) < 0$~~ can be also satisfied if α_j, b picked to be 0 and $\xi_i > 1$.

$g_{3,\bar{i}}(\hat{x})$, $g_{3,i}(\hat{x})$ can be < 0 by picking positive α_i and ξ_i.

=) therefore, by strong duality theorem, the duality gap is 0.

(Q6c).

$\nabla_{\alpha} L = \vec{1} - \sum_{i=1}^{N}\sum_{j=1}^{N} w_i y_i y_j x_i x_j - \sum_{i=1}^{N} \beta_i$

$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} w_i y_i$

$\frac{\partial L}{\partial \xi_i} = C_i - w_i - \gamma_i$

(*): $\sum_{i=1}^{N} w_i y_i = 0$, $w_i + \gamma_i = C_i$, $\forall i = 1, \cdots, N$

if (*) is satisfied, then $\Theta(w, \beta, \gamma) = L(\alpha, b, \xi, w, \beta, \gamma)$ iff $\vec{\alpha} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i y_i y_j x_i x_j + \sum_{j=1}^{N} \beta_j$

otherwise, one can make $\Theta(w, \beta, \gamma)$ to be $\Theta, -\infty$

So $\Theta(w, \beta, \gamma) = \frac{1}{2}\sum_{i=1}^{N} \alpha_i^2 + \sum_{i=1}^{N} C\xi_i + \sum_{i=1}^{N} w_i (1 - \xi_i - y_i(\sum_{j=1}^{N} a_j y_j x_i \cdot x_j + b)) + \sum_{i=1}^{N} \beta_i(-\alpha_i) + \sum_{i=1}^{N} \gamma_i(-\xi_i)$

$= \frac{1}{2}\sum_{i=1}^{N} \alpha_i^2 + \sum_{i=1}^{N} w_i - \alpha^T \underbrace{\left(\sum_{i=1}^{N}\sum_{j=1}^{N} w_i y_i y_j x_i \cdot x_j + \sum_{j=1}^{N} \beta_j\right)}_{\alpha}$

$= \sum_{i=1}^{N} w_i - \frac{1}{2}\|\alpha\|^2$

$= \sum_{i=1}^{N} w_i - \frac{1}{2}\left(\sum_{j=1}^{N}\sum_{j=1}^{N} w_i y_i y_j x_i \cdot x_j + \sum_{j=1}^{N} \beta_j\right)^T \left(\sum_{k=1}^{N}\sum_{l=1}^{N} w_k y_k y_l x_k \cdot x_l + \sum_{l=1}^{N} \beta_l\right)$ (continue)

d). ~~simplify~~ Simplifying the above, we have:

maximize $\Theta(w, \beta, \gamma) = \sum_{i=1}^{N} w_i - \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} w_i y_i y_j x_i \cdot x_j + \sum_{j=1}^{N} \beta_j\right)^T \cdot \left(\sum_{i=1}^{N}\sum_{j=1}^{N} w_i y_i y_j x_i \cdot x_j + \sum_{i=1}^{N} \beta_i\right)$

$(\Rightarrow)$ maximize $\Theta(w, \beta, \gamma) = \sum_{i=1}^{N} w_i - \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} y_j y_i x_j \cdot x_i\right)^2$

equivalent to

(as $\beta_i \geq 0$ is one of the condition in Lagrangian dual problem)

and also, the conditions $w_i + \gamma_i = C_i$, $w_i \geq 0$, $\gamma_i \geq 0$

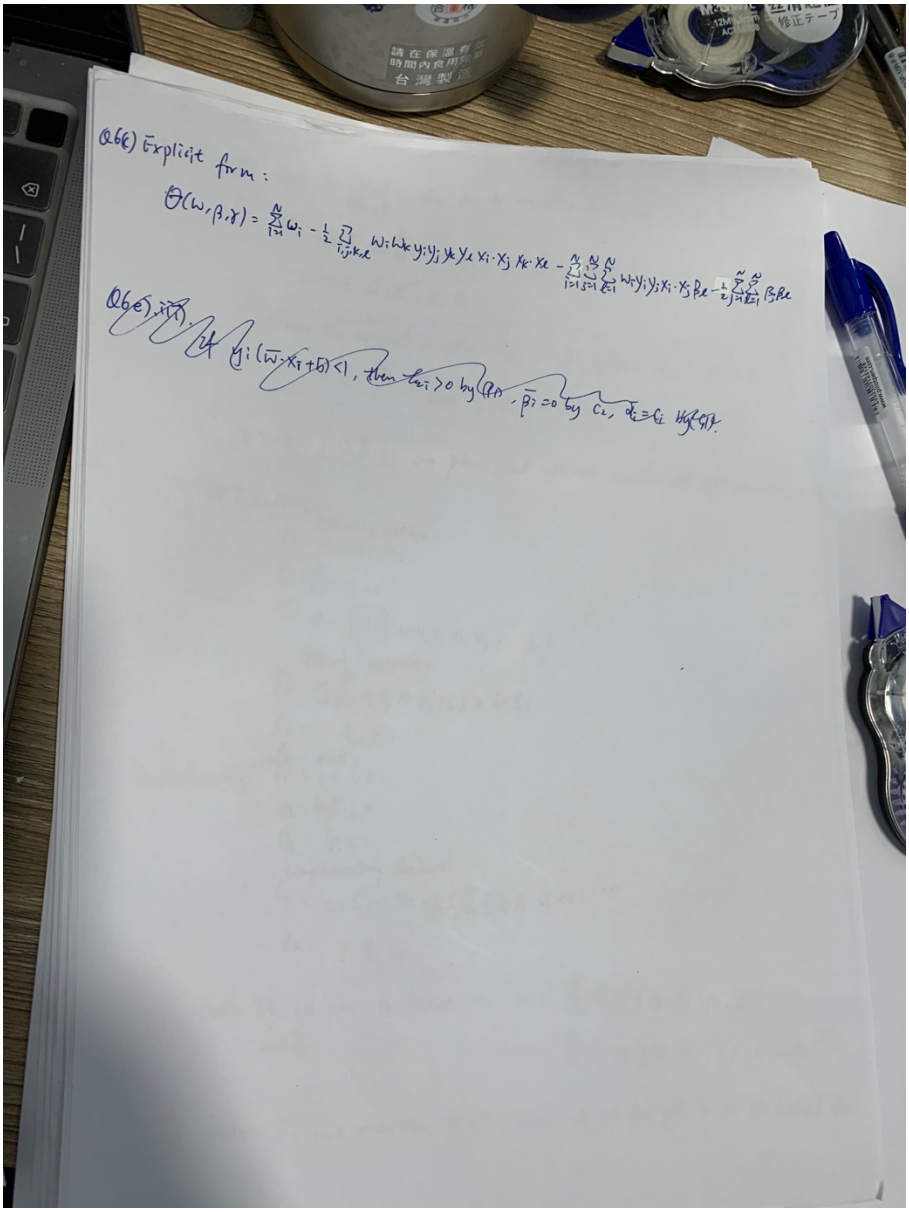
can be simplified to

$0 \leq w_i \leq C_i$, $\forall i = 1, \cdots, N$.

Therefore, the dual problem can be simplified as:

maximize $\sum_{i=1}^{N} w_i - \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} w_j y_j y_i x_j \cdot x_i\right)^2$

subject to $\sum_{i=1}^{N} w_i y_i = 0$, variables $0 \leq w_i \leq C_i$, $i = 1, \cdots N$



(b)(i) Explicit form:

$$\Theta(w,\beta,\gamma) = \sum_{i=1}^{N} w_i - \frac{1}{2}\sum_{i,j,k,l} w_i w_k y_i y_j y_k y_l x_i \cdot x_j x_k \cdot x_l - \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} w_i y_i y_j x_i \cdot x_j \beta_k - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\beta_j\beta_k$$

(b)(ii) If $y_i(\vec{w}\cdot\vec{x}_i+b)<1$, then $\xi_i>0$ by (ii), $\beta_i=0$ by C_2, $\alpha_i=C_1$ by (i).

Q6 e). i). Note that the ~~dual~~ primal problem can be rewritten as: (from rearranging terms in g_i and g_3 for t_i)

Minimize $\frac{1}{2}\|\alpha\|^2 + \sum_{i=1}^{N} C_i \max(1 - y_i(\sum_{j=1}^{N} \alpha_j y_j x_i \cdot x_j + b), 0)$

variables $\vec{\alpha} \in \mathbb{R}^N$, $b \in \mathbb{R}$.

Given the optimal weighting vector $\vec{w}$, the optimal bias is given by

$$b = \underset{b \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{N} C_i \max(1 - y_i(\vec{w} \cdot x_i + b), 0)$$

$(\vec{\alpha}, b, \vec{\xi})$, $(\vec{w}, \vec{\beta}, \vec{\gamma})$ are primal/dual optimal solutions iff KKT conditions hold

KKT conditions:

Stationary condition:

S_1: $w_i + \gamma_i = C_i$

S_2: $\sum_{i=1}^{N} w_i y_i = 0$

S_3: $\vec{x} = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} w_i y_i y_j x_i \cdot x_j + \beta_i \right)$

Primal feasibility:

P_1: $y_i(\sum_{j=1}^{N} \alpha_j y_j x_i \cdot x_j + b) \geq 1 - \xi_i$

P_2: $\xi_i \geq 0$

P_3: $\alpha_i \geq 0$

Dual feasibility → D_1: $w_i \geq 0$

D_2: $\beta_i \geq 0$

D_3: $\gamma_i \geq 0$

Complementary slackness

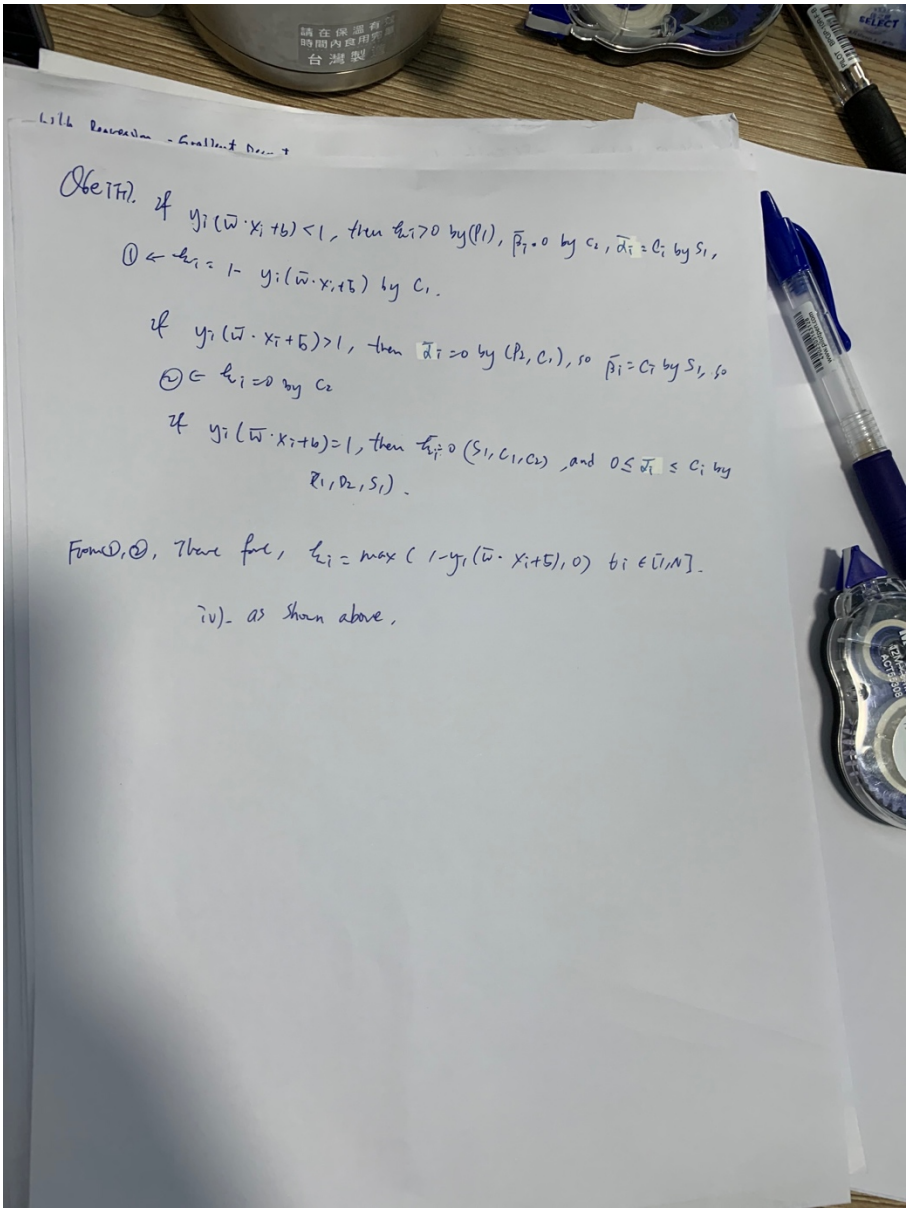
C_1: $w_i(1 - \xi_i - y_i(\sum_{j=1}^{N} \alpha_j y_j x_i \cdot x_j + b)) = 0$

C_2: $\gamma_i \xi_i = 0$

i). From S_3, we can see that $\alpha_i = \max(\sum_{j=1}^{N} w_j y_j y_i x_j \cdot x_i + \beta_i, 0)$

and P_3 $= \max(\sum_{j=1}^{N} w_j y_j y_i x_j \cdot x_i, 0)$ since $\beta_i \geq 0$ (D2).

ii). From P_1 and rewritten primal problem, or we can get b as the desired form



Linear Regression - Gradient Descent

Obs(Th). if $y_i(\vec{w}\cdot x_i+b)<1$, then $\xi_i>0$ by(P1), $\beta_i=0$ by C_2, $\alpha_i=C_i$ by S_1,

① ← $\xi_i = 1-y_i(\vec{w}\cdot x_i+b)$ by C_1.

if $y_i(\vec{w}\cdot x_i+b)>1$, then $\alpha_i=0$ by (P_2, C_1), so $\beta_i=C_i$ by S_1, so

② ← $\xi_i=0$ by C_2

if $y_i(\vec{w}\cdot x_i+b)=1$, then $\xi_i=0$ (S_1, C_1, C_2), and $0\le\alpha_i\le C_i$ by P_1, P_2, S_1).

From①,②, Therefore, $\xi_i = \max(1-y_i(\vec{w}\cdot x_i+b), 0)$ $\forall i\in[1,N]$.

iv). as shown above.

Q2 a). The joint probability function

$$= \prod_{i=1}^{n} p_\theta(\xi \mid x_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(\xi - f_\theta(x_i))^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (\xi - f_\theta(x_i))^2\right)$$

The log-likelihood function

$$\ell = \frac{-n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\xi - f_\theta(x_i))^2$$

Maximizing the log-likelihood is equivalent to minimizing

$$\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\xi - f_\theta(x_i))^2$$

or simply $\sum_{i=1}^{n} (\xi - f_\theta(x_i))^2$

$$\therefore \quad L_{ML}(\theta) = \sum_{i=1}^{n} (\xi - f_\theta(x_i))^2$$

b). $P(\theta \mid D_x) p(D_y \mid \theta, D_x) = \prod_{i=1}^{L} N(\theta^{(i)}; 0, \lambda^2) \prod_{i=1}^{N} N(y_i; f_\theta(x_i), \sigma^2)$

$$= \prod_{i=1}^{L} \frac{1}{\sqrt{2\pi\lambda}} \exp\left(\frac{-(\theta^{(i)} - 0)^2}{2\lambda^2}\right) \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(y_i - f_\theta(x))^2}{2\sigma^2}\right)$$

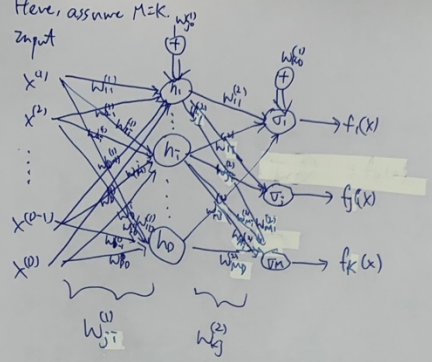
$$= \left(\frac{1}{\sqrt{2\pi\lambda^2}}\right)^{L} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{N} \exp\left(-\frac{1}{2\lambda^2} \sum_{i=1}^{L} (\theta^{(i)})^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2\right)$$

$$P(\theta \mid D) = \frac{P(\theta, D_y \mid D_x)}{P(D_y \mid D_x)} \propto P(\theta, D_y \mid D_x) = P(\theta \mid D_x)\, p(D_y \mid \theta, D_x)$$

$$\propto \exp(-L_{bayes}(\theta))$$

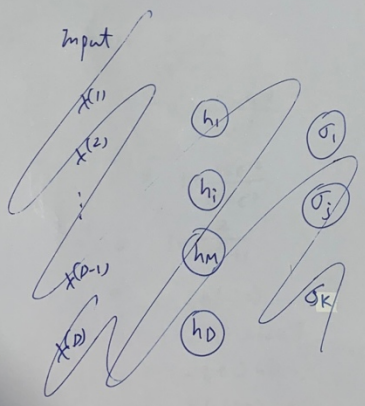
Hence proved.

Q(a). Here, assume M=K.

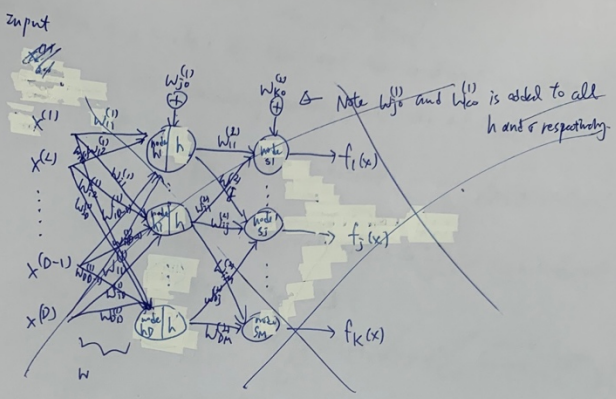


j denote index in x k denote index in 1~K

i denote index in h.



Q1 a).



b). $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

$= \frac{e^z + e^{-z} - 2e^{-z}}{e^z + e^{-z}}$

$= 1 + \frac{-2e^{-z}}{e^z + e^{-z}}$

$= 1 - \frac{2}{e^{2z} + 1}$

$= 1 - 2h(-2z)$

$= 1 - 2(1 - h(2z))$

$= 1 - 2 + 2h(2z)$

$= 2h(2z) - 1$

So, they differ by linear transformation. We can transform ~~the~~ $\tanh(z)$ by the above formula $2h(2z) - 1$, therefore there exist a ~~two~~ NN s.t. they perform the same function.

Q3: Our goal is to minimize L w.r.t. both α_t and $f_t(x)$.

We suppose $f_1, \dots, f_{t-1}$ are fixed, and their coefficients $\alpha_1, \dots, \alpha_{t-1}$ so we are minimizing only w.r.t. α_t and $f_t(x)$.

$$L = C_1 \sum_{i:y_i=1} e^{-g_{t-1}(x_i) - \alpha_t f_t(x_i)} + C_2 \sum_{i:y_i=-1} e^{g_{t-1}(x_i) - \alpha_t f_t(x_i)}$$

$$= C_1 \sum_{i:y_i=1} w_i^{(t)} \exp\{-\alpha_t f_t(x_i)\} + C_2 \sum_{i:y_i=-1} w_i^{(t)} \exp\{-\alpha_t f_t(x_i)\}$$

where $w_i^{(t)} = \exp\{-g_{t-1}(x_i)\}$

min ∂L w.r.t. α_t: $\frac{\partial L}{\partial \alpha_t} =$ $e^{-\alpha_t} \cdot \sum_{i=1}^{n} w_i^{(t)} \mathbb{I}(f_t(x_i) \neq y_i) - e^{\alpha_t} \sum_{i=1}^{n} w_i^{(t)} \mathbb{I}(f_t(x_i) \neq y_i)$

Solving the above equation, we have

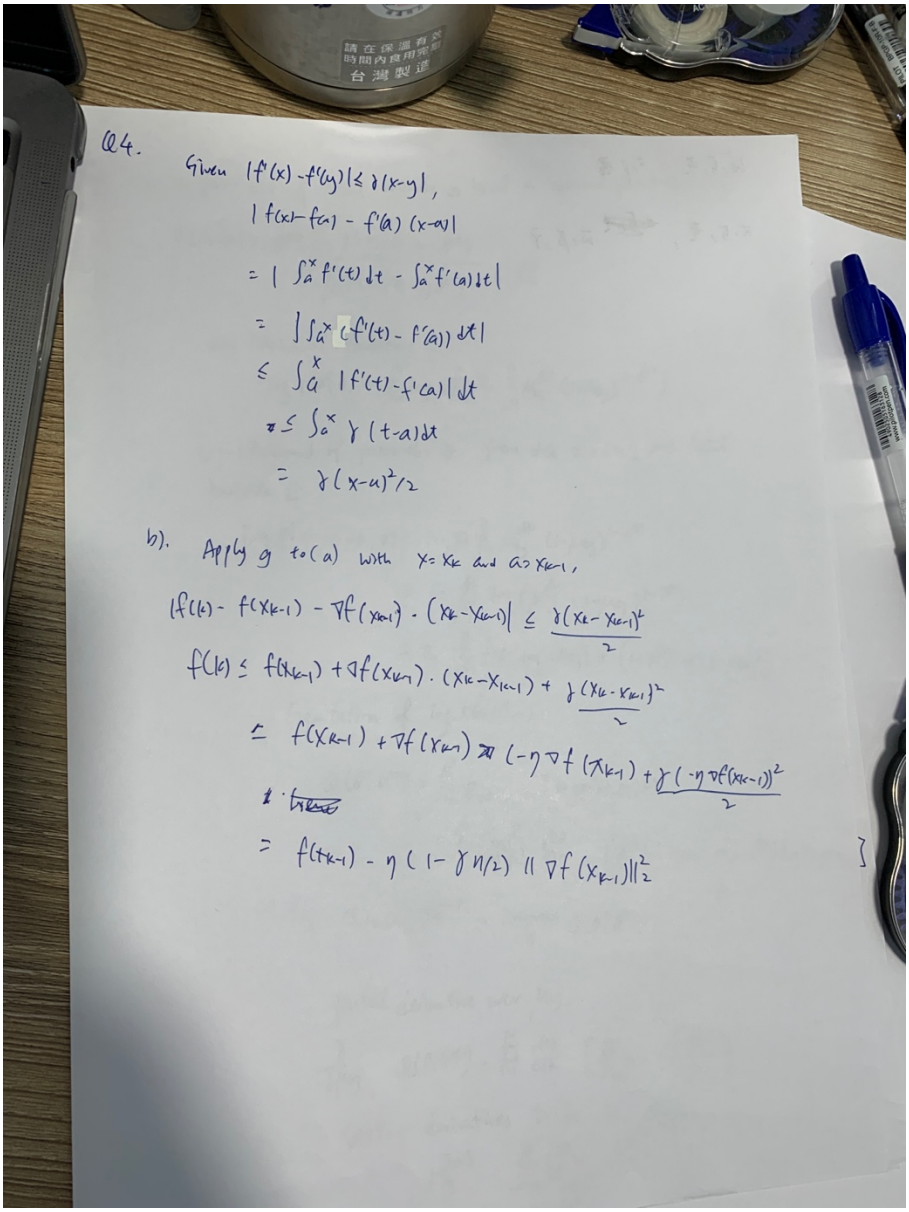
$\alpha_t = \log($

$$\frac{\partial L}{\partial \alpha_t} = -C_1 \sum_{i:y_i=1} w_i^{(t)} f_t(x_i) \exp\{-\alpha_t f_t(x_i)\} - C_2 \sum_{i:y_i=-1} w_i^{(t)} f_t(x_i) \exp\{-\alpha_t f_t(x_i)\} = 0$$

Solving above equation,

$$-C_1 \sum_{i:y_i=1} w_i^{(t)} f_t(x_i) = C_2 \sum_{i:y_i=-1} w_i^{(t)} f_t(x_i)$$

$$C_1 = \frac{-C_2 \sum_{i:y_i=-1} w_i^{(t)} f_t(x_i)}{\sum_{i:y_i=1} w_i^{(t)} f_t(x_i)}$$



Q4.

Given $|f'(x) - f'(y)| \le \gamma |x-y|$,

$$|f(x) - f(a) - f'(a)(x-a)|$$

$$= \left| \int_a^x f'(t)\,dt - \int_a^x f'(a)\,dt \right|$$

$$= \left| \int_a^x (f'(t) - f'(a))\,dt \right|$$

$$\le \int_a^x |f'(t) - f'(a)|\,dt$$

$$\le \int_a^x \gamma (t-a)\,dt$$

$$= \gamma (x-a)^2 / 2$$

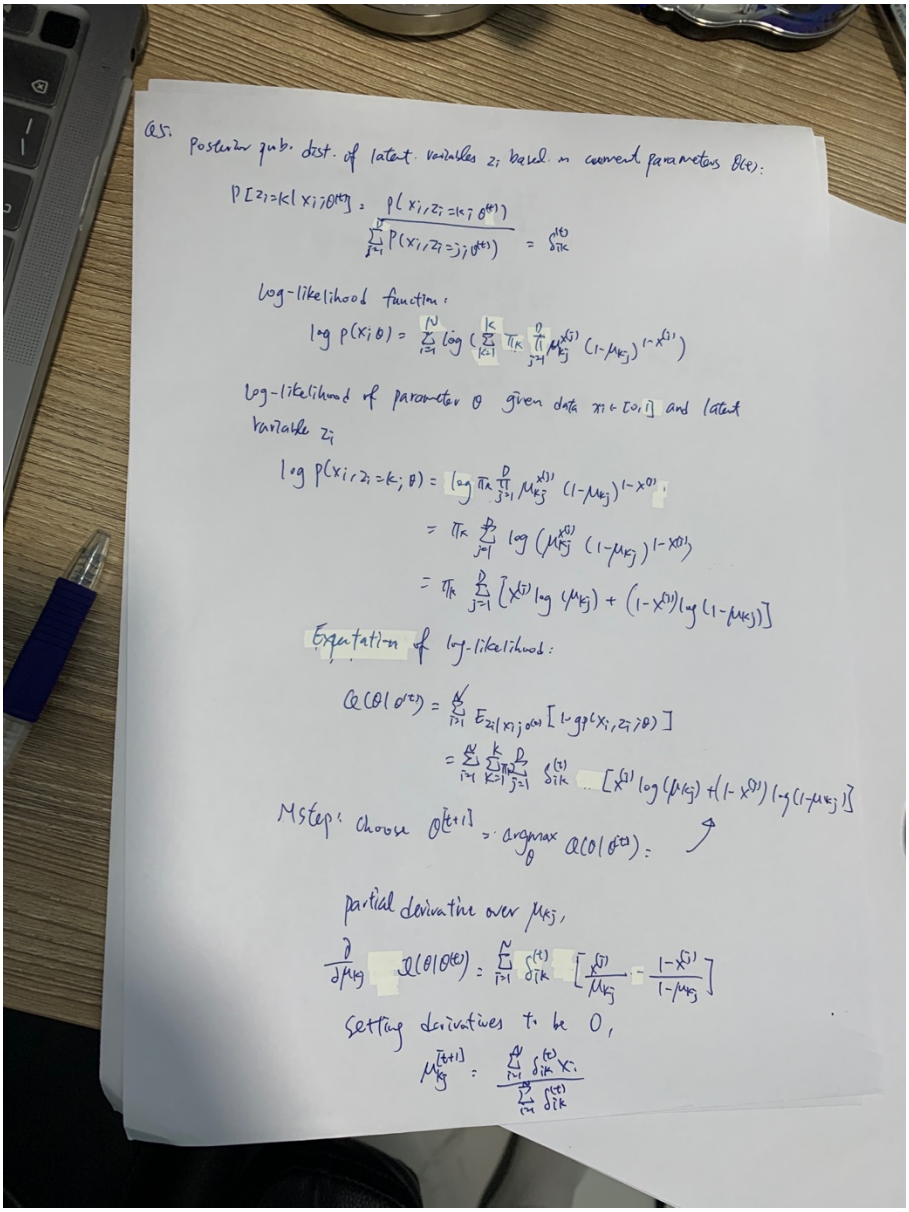
b). Apply g to (a) with $x = x_k$ and $a = x_{k-1}$,

$$|f(x_k) - f(x_{k-1}) - \nabla f(x_{k-1}) \cdot (x_k - x_{k-1})| \le \frac{\gamma (x_k - x_{k-1})^2}{2}$$

$$f(x_k) \le f(x_{k-1}) + \nabla f(x_{k-1}) \cdot (x_k - x_{k-1}) + \frac{\gamma (x_k - x_{k-1})^2}{2}$$

$$= f(x_{k-1}) + \nabla f(x_{k-1}) \cdot (-\eta \nabla f(x_{k-1})) + \frac{\gamma (-\eta \nabla f(x_{k-1}))^2}{2}$$

$$= f(x_{k-1}) - \eta \left(1 - \gamma \eta / 2\right) \|\nabla f(x_{k-1})\|_2^2$$



Q5.

Posterior prob. dist. of latent variables z_i based on current parameters $\theta^{(t)}$:

$$P[z_i=k \mid x_i;\theta^{(t)}] = \frac{p(x_i, z_i=k;\theta^{(t)})}{\sum_{j=1}^{K} P(x_i, z_i=j;\theta^{(t)})} = \delta_{ik}^{(t)}$$

log-likelihood function:

$$\log p(X;\theta) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \prod_{j=1}^{D} \mu_{kj}^{x_i^{(j)}} (1-\mu_{kj})^{1-x_i^{(j)}} \right)$$

log-likelihood of parameter θ given data $x_i \in [0,1]$ and latent variable z_i

$$\log p(x_i, z_i=k;\theta) = \log \pi_k \prod_{j=1}^{D} \mu_{kj}^{x_i^{(j)}} (1-\mu_{kj})^{1-x_i^{(j)}}$$

$$= \pi_k \sum_{j=1}^{D} \log \left(\mu_{kj}^{x_i^{(j)}} (1-\mu_{kj})^{1-x_i^{(j)}} \right)$$

$$= \pi_k \sum_{j=1}^{D} \left[x_i^{(j)} \log(\mu_{kj}) + (1-x_i^{(j)}) \log(1-\mu_{kj}) \right]$$

Expectation of log-likelihood:

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{N} E_{z_i \mid x_i;\theta^{(t)}} [\log p(x_i, z_i;\theta)]$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{D} \delta_{ik}^{(t)} \left[x_i^{(j)} \log(\mu_{kj}) + (1-x_i^{(j)}) \log(1-\mu_{kj}) \right]$$

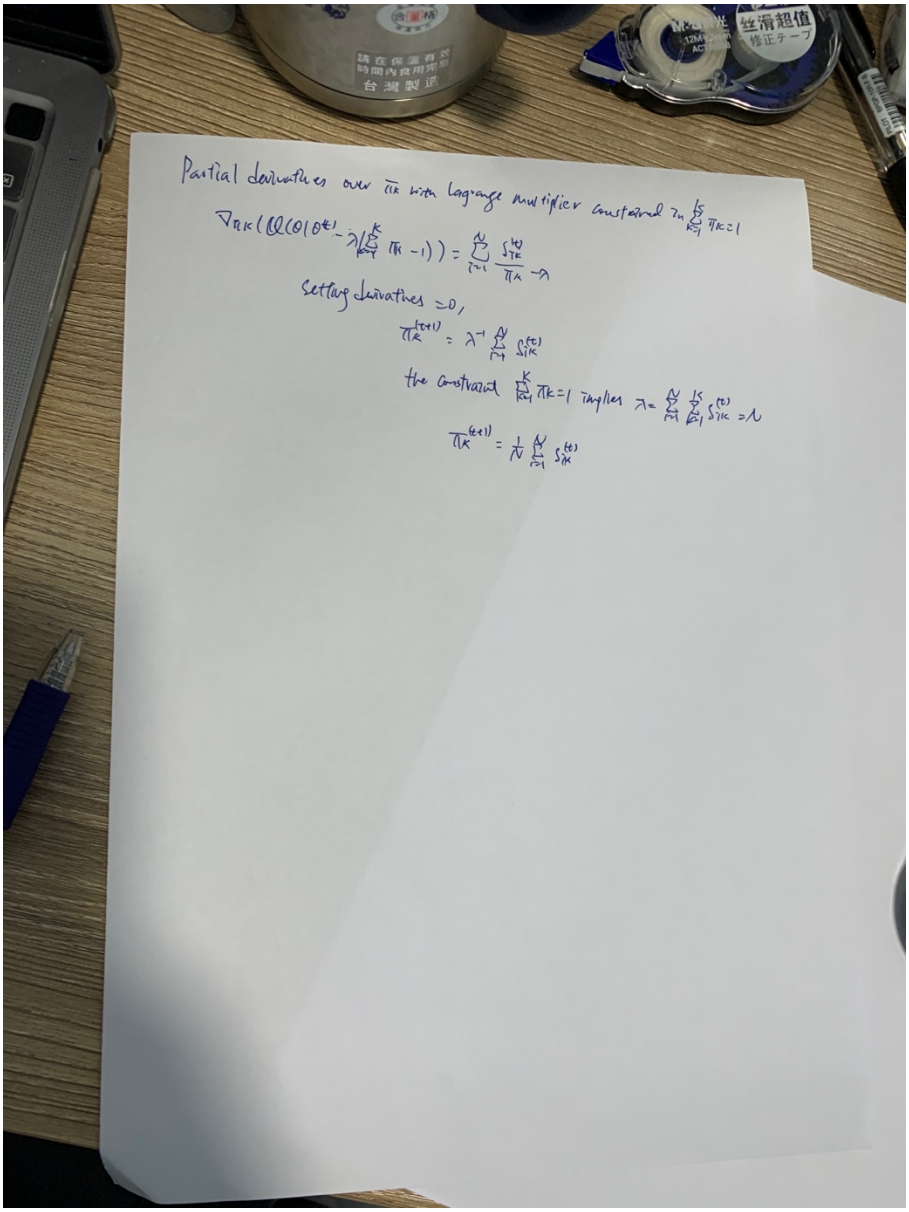
M step: choose $\theta^{(t+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(t)})$:

partial derivative over μ_{kj},

$$\frac{\partial}{\partial \mu_{kj}} Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{N} \delta_{ik}^{(t)} \left[\frac{x_i^{(j)}}{\mu_{kj}} - \frac{1-x_i^{(j)}}{1-\mu_{kj}} \right]$$

Setting derivatives to be 0,

$$\mu_{kj}^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} x_i}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$$



Partial derivatives over π_k with Lagrange multiplier constrained $\sum_{k=1}^{K} \pi_k = 1$

$$\nabla_{\pi_k}\left(Q(\theta|\theta^{(t)}) - \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)\right) = \sum_{i=1}^{N}\frac{s_{ik}^{(t)}}{\pi_k} - \lambda$$

Setting derivatives $= 0$,

$$\pi_k^{(t+1)} = \lambda^{-1}\sum_{i=1}^{N} s_{ik}^{(t)}$$

the constraint $\sum_{k=1}^{K}\pi_k = 1$ implies $\lambda = \sum_{i=1}^{N}\sum_{k=1}^{K} s_{ik}^{(t)} = N$

$$\pi_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} s_{ik}^{(t)}$$