HW4 Answer

Problem 1

t	1	2	3	4
z_i	90	90	190	90
$f(z_i)$	1	1	1	1
z_f	10	10	-90	10
$f(z_i)$	1	1	0	1
z_o	-10	90	90	90
$f(z_o)$	0	1	1	1
z	3	-2	4	0
c'	3	1	4	4
y	0	1	4	4

Problem 2

First, give an toy example: Consider a fully connected network with $\tanh$ as the non-linear activation e.g.

$$
\begin{gathered}
Y = WX + B \\
Z = \tanh(Y) \\
\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} w_1 & w_2 \\ w_3 & w_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\
\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \tanh(y_1) \\ \tanh(y_2) \end{bmatrix}
\end{gathered}
$$

If L is the loss of the network and given $\frac{\partial L}{\partial Z}$ (from the preceding layer), then

$$\begin{aligned}
\frac{\partial L}{\partial Y} &= \begin{bmatrix} \frac{\partial L}{\partial y_1} \\ \frac{\partial L}{\partial y_2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial L}{\partial z_1}\left(1 - z_1^2\right) \\ \frac{\partial L}{\partial z_2}\left(1 - z_2^2\right) \end{bmatrix} \\
&= \begin{bmatrix} \left(1 - z_1^2\right) \\ \left(1 - z_2^2\right) \end{bmatrix} \odot \begin{bmatrix} \frac{\partial L}{\partial z_1} \\ \frac{\partial L}{\partial z_2} \end{bmatrix} \\
&= \begin{bmatrix} \left(1 - z_1^2\right) & 0 \\ 0 & \left(1 - z_2^2\right) \end{bmatrix} \begin{bmatrix} \frac{\partial L}{\partial z_1} \\ \frac{\partial L}{\partial z_2} \end{bmatrix} \\
&= \operatorname{diag}\left(1 - Z^2\right) \frac{\partial L}{\partial Z} \\
\frac{\partial L}{\partial W} &= \frac{\partial L}{\partial Y} X^T \\
\frac{\partial L}{\partial X} &= W^T \frac{\partial L}{\partial Y}
\end{aligned}$$

To write explicity, let $W_1 = W_2 = W_h$ and $U_1 = U_2 = W_i$, then

$$\begin{aligned}
h_1 &= \tanh(W_1 h_0 + U_1 x_1) \\
h_2 &= \tanh(W_2 h_1 + U_2 x_2)
\end{aligned}$$

Let $z = W_o h_2$. By the chain rule, we can derive the results as follows:

$$\begin{aligned}
\frac{\partial L}{\partial W_o} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial W_o} \\
\frac{\partial L}{\partial W} &= \frac{\partial L}{\partial W_1} + \frac{\partial L}{\partial W_2} \\
&= \operatorname{diag}\left(1 - (h_1)^2\right) \frac{\partial L}{\partial h_1} h_0^T + \operatorname{diag}\left(1 - (h_2)^2\right) \frac{\partial L}{\partial h_2} h_1^T \\
\frac{\partial L}{\partial U} &= \frac{\partial L}{\partial U_1} + \frac{\partial L}{\partial U_2} \\
&= \operatorname{diag}\left(1 - (h_1)^2\right) \frac{\partial L}{\partial h_1} x_1^T + \operatorname{diag}\left(1 - (h_2)^2\right) \frac{\partial L}{\partial h_2} x_2^T
\end{aligned}$$

Note that $W_i, W_h \in \mathbb{R}^{n \times n}$, $h_2, h_1 \in \mathbb{R}^n$, and $W_o \in \mathbb{R}^{1 \times n}$. Let $L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log{(1 - \hat{y})}$

1. $\frac{\partial L(y,\hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z}$:
   $\frac{\partial L(y,\hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} = -[y \cdot \frac{1}{\hat{y}} + (1 - y) \cdot \frac{-1}{1-\hat{y}}] \cdot \hat{y}(1 - \hat{y}) = \hat{y} - y$
2. $\frac{\partial z}{\partial W_o}$:
   $\frac{\partial z}{\partial W_o} = h_2^T$
3. $\frac{\partial z}{\partial h_2}$: W_o^T
   By (1)(2), we can get $\frac{\partial L(y,\hat{y})}{\partial W_o} = h_2^T(\hat{y} - y)$
   Now,

$$\frac{\partial L}{\partial h_2} = W_o^T \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z}$$
$$\frac{\partial L}{\partial h_1} = W_h^T \operatorname{diag}\left(1 - (h_2)^2\right) \frac{\partial L}{\partial h_2}$$

Then, we can get $\frac{\partial L(y,\hat{y})}{\partial W_h} = (\hat{y} - y) diag(1 - (h_2)^2) W_o^T h_1^T$, and $\frac{\partial L(y,\hat{y})}{\partial W_i} = (\hat{y} - y) diag(1 - (h_1)^2) W_h^T diag(1 - (h_2)^2) W_o^T x_1^T + diag(1 - (h_2)^2) W_o^T x_2^T$.

Problem 3

Given $\mathbf{g}_{t-1} = \left\{g_{t-1}^k\right\}_{k=1}^K$, we update $\mathbf{g}_t = \left\{g_{t-1}^k + \alpha_t f_t^k\right\}_{k=1}^K = \mathbf{g}_{t-1} + \alpha_t \mathrm{f}_t$ as follows:

$$\begin{aligned}
&\mathbf{f}_t \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{\partial}{\partial \alpha} L\left(\mathbf{g}_{t-1} + \alpha \mathbf{f}\right)\Bigg|_{\alpha=0} \\
&= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{\partial}{\partial \alpha} \sum_{i=1}^n \exp\left(\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_{t-1}^k\left(x_i\right) - g_{t-1}^{\hat{y}_i}\left(x_i\right)\right) + \alpha\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} f^k\left(x_i\right) - f^{\hat{y}_i}\left(x_i\right)\right)\right)\Bigg|_{\alpha=0} \\
&= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^n \exp\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_{t-1}^k\left(x_i\right) - g_{t-1}^{\hat{y}_i}\left(x_i\right)\right)\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} f^k\left(x_i\right) - f^{\hat{y}_i}\left(x_i\right)\right) \\
&= \underset{f \in \mathcal{F}}{\operatorname{argmin}} Z_t \mathbb{E}_{i \sim D_t}\left[\frac{1}{K-1} \sum_{k \neq \hat{y}_i} f^k\left(x_i\right) - f^{\hat{y}_i}\left(x_i\right)\right] = \underset{f \in \mathcal{F}}{\operatorname{argmin}} Z_t \mathbb{E}_{i \sim D_t}\left[\frac{1}{K-1} \cdot 1\left\{f\left(x_i\right) \neq \hat{y}_i\right\} - 1\left\{f\left(x_i\right)\right.\right. \\
&= \underset{f \in \mathcal{F}}{\operatorname{argmin}} Z_t\left(\frac{K}{K-1} \mathbb{P}_{i \sim D_t}\left[f\left(x_i\right) \neq \hat{y}_i\right] - 1\right) = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \mathbb{P}_{i \sim D_t}\left[f\left(x_i\right) \neq \hat{y}_i\right] \\
&\alpha_t \in \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} L\left(\mathbf{g}_{t-1} + \alpha \mathbf{f}_t\right) \\
&= \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^n \exp\left(\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_{t-1}^k\left(x_i\right) - g_{t-1}^{\hat{y}_i}\left(x_i\right)\right) + \alpha\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} f_t^k\left(x_i\right) - f_t^{\hat{y}_i}\left(x_i\right)\right)\right) \\
&= \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} Z_t \mathbb{E}_{i \sim D_t}\left[e^{\alpha\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} f_t^k(x_i) - f_t^{\hat{y}_i}(x_i)\right)}\right] \\
&= \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} Z_t \mathbb{E}_{i \sim D_t}\left[e^{\frac{\alpha}{K-1}} \cdot 1\left\{f_t\left(x_i\right) \neq \hat{y}_i\right\} + e^{-\alpha} \cdot 1\left\{f_t\left(x_i\right) = \hat{y}_i\right\}\right] \\
&= \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} Z_t\left(\epsilon_t e^{\frac{\alpha}{K-1}} + e^{-\alpha}\left(1 - \epsilon_t\right)\right) = \left\{\frac{K-1}{K} \log \frac{(K-1)\left(1 - \epsilon_t\right)}{\epsilon_t}\right\}
\end{aligned}$$

where

$$Z_t = \sum_{i=1}^n \exp\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_{t-1}^k\left(x_i\right) - g_{t-1}^{\hat{y}_i}\left(x_i\right)\right)$$

and that D_t is a probability distribution for $t = 1, \cdots, n$ given by

$$D_t(i) = \frac{1}{Z_t} \exp\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_{t-1}^k(x_i) - g_{t-1}^{\hat{y}_i}(x_i) \right)$$

and that $\epsilon_t = \mathbb{P}_{i \sim D_t}[f_t(x_i) \neq \hat{y}_i]$ is the error of f_t on training sample weighted by the distribution D_t.

Problem 4

Follow the lecture note (https://ntueemlta2022.github.io/slides/week9/W9_GMM_EM.pdf (https://ntueemlta2022.github.io/slides/week9/W9_GMM_EM.pdf)). Calculate

$$\begin{aligned}
\pi_k^{(t+1)} &= \frac{1}{N} \sum_{i=1}^{N} \delta_{ik}^{(t)} \\
\boldsymbol{\mu}_k^{(t+1)} &= \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} \boldsymbol{x}_i}{\sum_{i=1}^{N} \delta_{ik}^{(t)}} \\
\boldsymbol{\Sigma}_k^{(t+1)} &= \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}
\end{aligned}$$

explicity to get all points.
The calculation process follows the HW2 answer. (之後"可能"會補上計算過程)