Department of Computer Science and Technology, Tsinghua University

Machine Learning HW2

Hongzun Liu 2022210866

October 27, 2022

1 Collaborators and Sources

I finished this assignment independently but refered to some blogs on the Internet.

References

1 Computing Neural Network Gradients - Stanford

2 Derivation of Batch Normalization's Gradient Backpropagation

3 Solving mixture of multinomial topic models with EM algorithm

2 Back Propagation

Problem 1

Useful Identities

Firstly, I will provide some useful identities which may be used in the computation process below.

Given $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$, we have

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \boldsymbol{W}$$

Given $\boldsymbol{z} = f(\boldsymbol{x}), \boldsymbol{\delta} = \frac{\partial L}{\partial \boldsymbol{z}}$, where $f(\cdot)$ is an element-wise funciton, we have

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = diag(f'(\boldsymbol{x}))$$

$$\frac{\partial L}{\partial \boldsymbol{x}} = \frac{\partial L}{\partial \boldsymbol{z}}\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = f'(\boldsymbol{x}) \odot \boldsymbol{\delta}$$

where $\odot$ is hadamand product (element-wise product).

Given $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}, \boldsymbol{\delta} = \frac{\partial L}{\partial \boldsymbol{z}}$, we have

$$\frac{\partial L}{\partial \boldsymbol{W}} = \frac{\partial L}{\partial \boldsymbol{z}}\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{W}} = \boldsymbol{\delta}\boldsymbol{x}^T$$

$$\frac{\partial L}{\partial \boldsymbol{x}} = \frac{\partial L}{\partial \boldsymbol{z}}\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \boldsymbol{W}^T\frac{\partial L}{\partial \boldsymbol{z}}$$

Preparation

For simplicity, we let $L = f_{CE}$. By using the Chain-rule, we have

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{z}_{2,i}} &= \frac{\partial L}{\partial \hat{\boldsymbol{y}}_i} \frac{\partial \hat{\boldsymbol{y}}_i}{\partial \boldsymbol{z}_{2,i}} \\
\frac{\partial L}{\partial \boldsymbol{W}^{(2)}} &= \sum_{i=1}^{n} \frac{\partial L}{\partial \boldsymbol{z}_{2,i}} \frac{\partial \boldsymbol{z}_{2,i}}{\partial \boldsymbol{W}^{(2)}} \\
\frac{\partial L}{\partial \boldsymbol{b}^{(2)}} &= \sum_{i=1}^{n} \frac{\partial L}{\partial \boldsymbol{z}_{2,i}} \frac{\partial \boldsymbol{z}_{2,i}}{\partial \boldsymbol{b}^{(2)}} \\
\frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} &= \frac{\partial L}{\partial \boldsymbol{z}_{2,i}} \frac{\partial \boldsymbol{z}_{2,i}}{\partial \hat{\boldsymbol{h}}_{1,i}} \\
\frac{\partial L}{\partial \gamma} &= \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} \frac{\partial \hat{\boldsymbol{h}}_{1,i}}{\partial \gamma} \\
\frac{\partial L}{\partial \beta} &= \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} \frac{\partial \hat{\boldsymbol{h}}_{1,i}}{\partial \beta} \\
\frac{\partial L}{\partial \boldsymbol{\sigma}^2} &= \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} \frac{\partial \hat{\boldsymbol{h}}_{1,i}}{\partial \boldsymbol{\sigma}^2} \\
\frac{\partial L}{\partial \boldsymbol{\mu}} &= \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} \frac{\partial \hat{\boldsymbol{h}}_{1,i}}{\partial \boldsymbol{\mu}} + \frac{\partial L}{\partial \boldsymbol{\sigma}^2} \frac{\partial \boldsymbol{\sigma}^2}{\partial \boldsymbol{\mu}} \\
\frac{\partial L}{\partial \boldsymbol{h}_{1,i}} &= \frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} \frac{\partial \hat{\boldsymbol{h}}_{1,i}}{\partial \boldsymbol{h}_{1,i}} + \frac{\partial L}{\partial \boldsymbol{\mu}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{h}_{1,i}} + \frac{\partial L}{\partial \boldsymbol{\sigma}^2} \frac{\partial \boldsymbol{\sigma}^2}{\partial \boldsymbol{h}_{1,i}} \\
\frac{\partial L}{\partial \boldsymbol{z}_{1,i}} &= \frac{\partial L}{\partial \boldsymbol{h}_{1,i}} \frac{\partial \boldsymbol{h}_{1,i}}{\partial \boldsymbol{z}_{1,i}} \\
\frac{\partial L}{\partial \boldsymbol{W}^{(1)}} &= \sum_{i=1}^{n} \frac{\partial L}{\partial \boldsymbol{z}_{1,i}} \frac{\partial \boldsymbol{z}_{1,i}}{\partial \boldsymbol{W}^{(1)}} \\
\frac{\partial L}{\partial \boldsymbol{b}^{(1)}} &= \sum_{i=1}^{n} \frac{\partial L}{\partial \boldsymbol{z}_{1,i}} \frac{\partial \boldsymbol{z}_{1,i}}{\partial \boldsymbol{b}^{(1)}}
\end{aligned}$$

Derivation

More specifically, we have

$$\begin{aligned}
\frac{\partial L}{\partial z_{2,i,k}} &= \sum_{j=1}^{n_y} \frac{\partial L}{\partial \hat{y}_{ij}} \frac{\partial \hat{y}_{ij}}{\partial z_{2,i,k}} \\
&= -\frac{1}{n} \left[\sum_{j=1, j\neq k}^{n_y} \frac{y_{ij}}{\hat{y}_{ij}} (-\hat{y}_{ik}\hat{y}_{ij}) + \frac{y_{ik}}{\hat{y}_{ik}} \hat{y}_{ik}(1-\hat{y}_{ik}) \right] \qquad (1) \\
&= \frac{1}{n} \left[\hat{y}_{ik}(\sum_{j=1}^{n_y} y_{ij}) - y_{ik} \right]
\end{aligned}$$

Noted that in practice, $\sum_{j=1}^{n_y} y_{ij}$ usally equals to 1.

Then

$$\boldsymbol{\delta}_{1,i} = \frac{\partial L}{\partial \boldsymbol{z}_{2,i}} = \frac{1}{n} \left(\left(\sum_{j=1}^{n_y} y_{ij} \right) \hat{\boldsymbol{y}}_i - \boldsymbol{y}_i \right) \in \mathbb{R}^{n_y} \qquad (2)$$

Thus

$$\frac{\partial L}{\partial \boldsymbol{W}^{(2)}} = \sum_{i=1}^{n} \frac{\partial L}{\partial \boldsymbol{z}_{2,i}} \frac{\partial \boldsymbol{z}_{2,i}}{\partial \boldsymbol{W}^{(2)}} = \sum_{i=1}^{n} \boldsymbol{\delta}_{1,i} \boldsymbol{h}_{1,i}^T \in \mathbb{R}^{n_y \times n_1} \qquad (3)$$

$$\frac{\partial L}{\partial \boldsymbol{b}^{(2)}} = \sum_{i=1}^{n} \frac{\partial L}{\partial \boldsymbol{z}_{2,i}} \frac{\partial \boldsymbol{z}_{2,i}}{\partial \boldsymbol{b}^{(2)}} = \sum_{i=1}^{n} \boldsymbol{\delta}_{1,i} \boldsymbol{I} = \sum_{i=1}^{n} \boldsymbol{\delta}_{1,i} \in \mathbb{R}^{n_y} \tag{4}$$

$$\delta_{2,i} = \frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} = \frac{\partial L}{\partial \boldsymbol{z}_{2,i}} \frac{\partial \boldsymbol{z}_{2,i}}{\partial \hat{\boldsymbol{h}}_{1,i}} = \boldsymbol{W}^{(2)^T} \boldsymbol{\delta}_{1,i} \in \mathbb{R}^{n_1} \tag{5}$$

$$\frac{\partial L}{\partial \gamma} = \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} \frac{\partial \hat{\boldsymbol{h}}_{1,i}}{\partial \gamma} = \sum_{i=1}^{n} \left(\frac{\boldsymbol{h}_{1,i} - \boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon}}} \right)^T \boldsymbol{\delta}_{2,i} \in \mathbb{R} \tag{6}$$

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} \frac{\partial \hat{\boldsymbol{h}}_{1,i}}{\partial \beta} = \sum_{i=1}^{n} \boldsymbol{1}^T \boldsymbol{\delta}_{2,i} \in \mathbb{R} \tag{7}$$

$$\frac{\partial L}{\partial \boldsymbol{\sigma}^2} = \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} \frac{\partial \hat{\boldsymbol{h}}_{1,i}}{\partial \boldsymbol{\sigma}^2} = -\frac{\gamma}{2} \sum_{i=1}^{n} \left(\frac{(\boldsymbol{h}_{1,i} - \boldsymbol{\mu})}{(\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon})^{\frac{3}{2}}} \odot \boldsymbol{\delta}_{2,i} \right) \in \mathbb{R}^{n_1} \tag{8}$$

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\mu}} &= \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} \frac{\partial \hat{\boldsymbol{h}}_{1,i}}{\partial \boldsymbol{\mu}} + \frac{\partial L}{\partial \boldsymbol{\sigma}^2} \frac{\partial \boldsymbol{\sigma}^2}{\partial \boldsymbol{\mu}} \\ &= -\sum_{i=1}^{n} \frac{\gamma}{\sqrt{\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon}}} \odot \boldsymbol{\delta}_{2,i} + \left(\frac{\gamma}{2} \sum_{i=1}^{n} \left(\frac{\boldsymbol{h}_{1,i} - \boldsymbol{\mu}}{(\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon})^{\frac{3}{2}}} \odot \boldsymbol{\delta}_{2,i} \right) \right) \odot \frac{2}{n} \sum_{i=1}^{n} (\boldsymbol{h}_{1,i} - \boldsymbol{\mu}) \\ &= -\sum_{i=1}^{n} \frac{\gamma}{\sqrt{\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon}}} \odot \boldsymbol{\delta}_{2,i} \\ &= -\frac{\gamma}{\sqrt{\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon}}} \odot \sum_{i=1}^{n} \boldsymbol{\delta}_{2,i} \in \mathbb{R}^{n_1} \end{aligned} \tag{9}$$

$$\begin{aligned} \boldsymbol{\delta}_{3,i} &= \frac{\partial L}{\partial \boldsymbol{h}_{1,i}} = \frac{\partial L}{\partial \hat{\boldsymbol{h}}_{1,i}} \frac{\partial \hat{\boldsymbol{h}}_{1,i}}{\partial \boldsymbol{h}_{1,i}} + \frac{\partial L}{\partial \boldsymbol{\mu}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{h}_{1,i}} + \frac{\partial L}{\partial \boldsymbol{\sigma}^2} \frac{\partial \boldsymbol{\sigma}^2}{\partial \boldsymbol{h}_{1,i}} \\ &= \frac{\gamma}{\sqrt{\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon}}} \odot \boldsymbol{\delta}_{2,i} - \left[\frac{\gamma}{\sqrt{\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon}}} \odot \sum_{j=1}^{n} \boldsymbol{\delta}_{2,j} \right] \cdot \frac{1}{n} - \left[\frac{\gamma}{2} \sum_{j=1}^{n} \left(\frac{(\boldsymbol{h}_{1,j} - \boldsymbol{\mu})}{(\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon})^{\frac{3}{2}}} \odot \boldsymbol{\delta}_{2,j} \right) \right] \odot \frac{2}{n} (\boldsymbol{h}_{1,i} - \boldsymbol{\mu}) \\ &= \frac{\gamma}{\sqrt{\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon}}} \odot \boldsymbol{\delta}_{2,i} - \frac{\gamma}{n} \left[\frac{1}{\sqrt{\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon}}} \odot \sum_{j=1}^{n} \boldsymbol{\delta}_{2,j} \right] - \frac{\gamma}{n} \left[\sum_{j=1}^{n} \left(\frac{(\boldsymbol{h}_{1,j} - \boldsymbol{\mu})}{(\boldsymbol{\sigma}^2 + \boldsymbol{\epsilon})^{\frac{3}{2}}} \odot \boldsymbol{\delta}_{2,j} \right) \right] \odot (\boldsymbol{h}_{1,i} - \boldsymbol{\mu}) \in \mathbb{R}^{n_1} \end{aligned} \tag{10}$$

$$\boldsymbol{\delta}_{4,i} = \frac{\partial L}{\partial \boldsymbol{z}_{1,i}} = \frac{\partial L}{\partial \boldsymbol{h}_{1,i}} \frac{\partial \boldsymbol{h}_{1,i}}{\partial \boldsymbol{z}_{1,i}} = \boldsymbol{1}\{\boldsymbol{z}_{1,i} > 0\} \odot \boldsymbol{\delta}_{3,i} \in \mathbb{R}^{n_1} \tag{11}$$

$$\frac{\partial L}{\partial \boldsymbol{W}^{(1)}} = \sum_{i=1}^{n} \frac{\partial L}{\partial \boldsymbol{z}_{1,i}} \frac{\partial \boldsymbol{z}_{1,i}}{\partial \boldsymbol{W}^{(1)}} = \sum_{i=1}^{n} \boldsymbol{\delta}_{4,i} \boldsymbol{x}_i^T \in \mathbb{R}^{n_1 \times n_x} \tag{12}$$

$$\frac{\partial L}{\partial \boldsymbol{b}^{(1)}} = \sum_{i=1}^{n} \frac{\partial L}{\partial \boldsymbol{z}_{1,i}} \frac{\partial \boldsymbol{z}_{1,i}}{\partial \boldsymbol{b}^{(1)}} = \sum_{i=1}^{n} \boldsymbol{\delta}_{4,i} \boldsymbol{1} = \sum_{i=1}^{n} \boldsymbol{\delta}_{4,i} \in \mathbb{R}^{n_1} \tag{13}$$

where $\odot$ is hadamand product (element-wise product), and $\boldsymbol{1}\{condition\}$ is a vector of which each element is 1 when *condition* is *true* and 0 otherwise.

In conclusion

$$
\begin{aligned}
\boldsymbol{\delta}_{1,i} = \frac{\partial f_{CE}}{\partial \boldsymbol{z}_{2,i}} &= \frac{1}{n}\left(\left(\sum_{j=1}^{n_y} y_{ij}\right)\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\right) && \in \mathbb{R}^{n_y} && (14)\\
\frac{\partial f_{CE}}{\partial \boldsymbol{W}^{(2)}} &= \sum_{i=1}^{n} \boldsymbol{\delta}_{1,i}\boldsymbol{h}_{1,i}^T && \in \mathbb{R}^{n_y \times n_1} && (15)\\
\frac{\partial f_{CE}}{\partial \boldsymbol{b}^{(2)}} &= \sum_{i=1}^{n} \boldsymbol{\delta}_{1,i} && \in \mathbb{R}^{n_y} && (16)\\
\boldsymbol{\delta}_{2,i} = \frac{\partial f_{CE}}{\partial \hat{\boldsymbol{h}}_{1,i}} &= {\boldsymbol{W}^{(2)}}^T \boldsymbol{\delta}_{1,i} && \in \mathbb{R}^{n_1} && (17)\\
\frac{\partial f_{CE}}{\partial \gamma} &= \sum_{i=1}^{n}\left(\frac{\boldsymbol{h}_{1,i}-\boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2+\boldsymbol{\epsilon}}}\right)^T \boldsymbol{\delta}_{2,i} && \in \mathbb{R} && (18)\\
\frac{\partial f_{CE}}{\partial \beta} &= \sum_{i=1}^{n} \mathbf{1}^T \boldsymbol{\delta}_{2,i} && \in \mathbb{R} && (19)\\
\boldsymbol{\delta}_{3,i} = \frac{\partial f_{CE}}{\partial \boldsymbol{h}_{1,i}} &= \frac{\gamma}{\sqrt{\boldsymbol{\sigma}^2+\boldsymbol{\epsilon}}} \odot \boldsymbol{\delta}_{2,i} - \frac{\gamma}{n}\left[\frac{1}{\sqrt{\boldsymbol{\sigma}^2+\boldsymbol{\epsilon}}} \odot \sum_{j=1}^{n}\boldsymbol{\delta}_{2,j}\right] \\
&\quad - \frac{\gamma}{n}\left[\sum_{j=1}^{n}\left(\frac{(\boldsymbol{h}_{1,j}-\boldsymbol{\mu})}{(\boldsymbol{\sigma}^2+\boldsymbol{\epsilon})^{\frac{3}{2}}} \odot \boldsymbol{\delta}_{2,j}\right)\right] \odot (\boldsymbol{h}_{1,i}-\boldsymbol{\mu}) && \in \mathbb{R}^{n_1} && (20)\\
\boldsymbol{\delta}_{4,i} = \frac{\partial f_{CE}}{\partial \boldsymbol{z}_{1,i}} &= \mathbf{1}\{\boldsymbol{z}_{1,i} > 0\} \odot \boldsymbol{\delta}_{3,i} && \in \mathbb{R}^{n_1} && (21)\\
\frac{\partial f_{CE}}{\partial \boldsymbol{W}^{(1)}} &= \sum_{i=1}^{n} \boldsymbol{\delta}_{4,i}\boldsymbol{x}_i^T && \in \mathbb{R}^{n_1 \times n_x} && (22)\\
\frac{\partial f_{CE}}{\partial \boldsymbol{b}^{(1)}} &= \sum_{i=1}^{n} \boldsymbol{\delta}_{4,i} && \in \mathbb{R}^{n_1} && (23)
\end{aligned}
$$

3 Mixtures of Logistic Models

Problem 2

1. Proof.

Let $\boldsymbol{\rho}_n = (\rho_{n1}, \rho_{n2}, \cdots, \rho_{nk})^T \in \mathbb{R}^K$, where $\rho_{nk} = y_{nk}^{t_n}(1-y_{nk})^{1-t_n}$. Thus

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k y_{nk}^{t_n}(1-y_{nk})^{1-t_n}\right) \\
&= \sum_{n=1}^{N} \log(\boldsymbol{\pi}^T \boldsymbol{\rho}_n)
\end{aligned} \tag{24}
$$

where $y_{nk} = \sigma(\boldsymbol{w}_k^T \boldsymbol{\phi}_n)$, $\boldsymbol{\pi} = (\pi_{n1}, \pi_{n2}, \cdots, \pi_{nk})^T \in \mathbb{R}^K$.

By using Lagrange multiplier, we will get

$$
\begin{aligned}
L(\boldsymbol{\pi}; \lambda) &= \sum_{n=1}^{N} \log(\boldsymbol{\pi}^T \boldsymbol{\rho}_n) - \lambda(\boldsymbol{\pi}^T \mathbf{1} - 1) \\
s.t. \quad & \nabla_{\boldsymbol{\pi}} L(\boldsymbol{\pi}; \lambda) = 0 \\
& \boldsymbol{\pi}^T \mathbf{1} = 1
\end{aligned} \tag{25}
$$

The optimal solution satisfies

$$\nabla_{\boldsymbol{\pi}} L(\boldsymbol{\pi};\lambda) = \sum_{n=1}^{N} \frac{\boldsymbol{\rho}_n}{\boldsymbol{\pi}^T \boldsymbol{\rho}_n} - \lambda \mathbf{1} = 0 \tag{26}$$

so we have

$$\sum_{n=1}^{N} \frac{\rho_{nk}}{\boldsymbol{\pi}^T \boldsymbol{\rho}_n} = \lambda, \quad k = 1, 2, \cdots, K \tag{27}$$

Let

$$\gamma_{nk} = \frac{\pi_k \rho_{nk}}{\boldsymbol{\pi}^T \boldsymbol{\rho}_n} = \frac{\pi_k y_{nk}^{t_n} (1-y_{nk})^{1-t_n}}{\sum_{j=1}^{K} \pi_j y_{nj}^{t_n} (1-y_{nj})^{1-t_n}} \tag{28}$$

Thus

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{\lambda} \tag{29}$$

We let $\boldsymbol{\pi}^T$ left multiplication equation 26, thus

$$\lambda = \lambda \sum_{k=1}^{K} \pi_k = \boldsymbol{\pi}^T \cdot \lambda \mathbf{1} = \sum_{n=1}^{N} \frac{\boldsymbol{\pi}^T \boldsymbol{\rho}_n}{\boldsymbol{\pi}^T \boldsymbol{\rho}_n} = N \tag{30}$$

Thus, we have

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N} \tag{31}$$

where $\gamma_{nk} = \dfrac{\pi_k y_{nk}^{t_n} (1-y_{nk})^{1-t_n}}{\sum_{j=1}^{K} \pi_j y_{nj}^{t_n} (1-y_{nj})^{1-t_n}}$. □

2. Proof.

Firstly,

$$\frac{\partial y_{nk}}{\partial \boldsymbol{w}_k} = \frac{\partial \sigma(\boldsymbol{w}_k^T \boldsymbol{\phi}_n)}{\partial \boldsymbol{w}_k} = y_{nk}(1-y_{nk}) \boldsymbol{\phi}_n \tag{32}$$

Then

$$\begin{aligned}
& \frac{\partial \left(\sum_{k=1}^{K} \pi_k y_{nk}^{t_n} (1-y_{nk})^{1-t_n} \right)}{\partial \boldsymbol{w}_k} \\
= \;& \frac{\partial \left(\pi_k y_{nk}^{t_n} (1-y_{nk})^{1-t_n} \right)}{\partial \boldsymbol{w}_k} \\
= \;& \pi_k t_n y_{nk} (1-y_{nk})^{1-t_n} \frac{\partial y_{nk}}{\partial \boldsymbol{w}_k} - (1-t_n)(1-y_{nk})^{-t_n} y_{nk}^{t_n} \frac{\partial y_{nk}}{\partial \boldsymbol{w}_k} \\
= \;& \pi_k \frac{y_{nk}^{t_n - 1}}{(1-y_{nk})^{t_n}} (t_n - y_{nk}) \frac{\partial \sigma(\boldsymbol{w}_k^T \boldsymbol{\phi}_n)}{\partial \boldsymbol{w}_k} \\
= \;& \pi_k y_{nk}^{t_n} (1-y_{nk})^{1-t_n} (t_n - y_{nk}) \boldsymbol{\phi}_n
\end{aligned} \tag{33}$$

Thus

$$\begin{aligned}
\nabla_{\boldsymbol{w}_k} L(\boldsymbol{w}_k) &= \sum_{n=1}^{N} \nabla_{\boldsymbol{w}_k} \log(\boldsymbol{\pi}^T \boldsymbol{\rho}_n) \\
&= \sum_{n=1}^{N} \nabla_{\boldsymbol{w}_k} \log \left(\sum_{k=1}^{K} \pi_k y_{nk}^{t_n} (1-y_{nk})^{1-t_n} \right) \\
&= \sum_{n=1}^{N} \frac{\pi_k y_{nk}^{t_n} (1-y_{nk})^{1-t_n}}{\sum_{k=1}^{K} \pi_k y_{nk}^{t_n} (1-y_{nk})^{1-t_n}} (t_n - y_{nk}) \boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N} \gamma_{nk} (t_n - y_{nk}) \boldsymbol{\phi}_n
\end{aligned} \tag{34}$$

$\square$

3. Proof.

$$
\begin{aligned}
\nabla_{\boldsymbol{w}_k}\left(\gamma_{nk}(t_n - y_{nk})\right)c &= \left(\nabla_{\boldsymbol{w}_k}\gamma_{nk}\right)(t_n - y_{nk}) - \gamma_{nk}\left(\nabla_{\boldsymbol{w}_k} y_{nk}\right) \\
&= \frac{\nabla_{\boldsymbol{w}_k}(\pi_k\rho_{nk}) \cdot \boldsymbol{\pi}^T\boldsymbol{\rho}_n - \nabla_{\boldsymbol{w}_k}(\pi_k\rho_{nk}) \cdot \pi_k\rho_{nk}}{(\boldsymbol{\pi}^T\boldsymbol{\rho}_n)^2}(t_n - y_{nk}) - \gamma_{nk}y_{nk}(1 - y_{nk})\boldsymbol{\phi}_n \\
&= \frac{1 - \gamma_{nk}}{\boldsymbol{\pi}^T\boldsymbol{\rho}_n}\nabla_{\boldsymbol{w}_k}(\pi_k\rho_{nk})(t_n - y_{nk}) - \gamma_{nk}y_{nk}(1 - y_{nk})\boldsymbol{\phi}_n \\
&= (1 - \gamma_{nk})\gamma_{nk}(t_n - y_{nk})^2\boldsymbol{\phi}_n - \gamma_{nk}y_{nk}(1 - y_{nk})\boldsymbol{\phi}_n \\
&= \left((1 - \gamma_{nk})\gamma_{nk}(t_n - y_{nk})^2 - \gamma_{nk}y_{nk}(1 - y_{nk})\right)\boldsymbol{\phi}_n
\end{aligned}
\tag{35}
$$

Hence

$$
\begin{aligned}
\boldsymbol{H}_k &= -\nabla_{\boldsymbol{w}_k}\left(\sum_{n=1}^{N}\gamma_{nk}(t_n - y_{nk})\boldsymbol{\phi}_n\right) \\
&= -\sum_{n=1}^{N}\nabla_{\boldsymbol{w}_k}\left(\gamma_{nk}(t_n - y_{nk})\right)\boldsymbol{\phi}_n^T \\
&= -\sum_{n=1}^{N}\left((1 - \gamma_{nk})\gamma_{nk}(t_n - y_{nk})^2 - \gamma_{nk}y_{nk}(1 - y_{nk})\right)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T
\end{aligned}
\tag{36}
$$

$\square$

4 Generative Adversarial Networks

Problem 3

Proof.
Obviously,

$$
p(z)dz = p_{model}(x)dx
$$

Thus

$$
\begin{aligned}
&\mathbb{E}_{p_{data}(x)}[\log D(x)] + \mathbb{E}_{p(z)}[\log(1 - D(G(z)))] \\
=& \int_x p_{data}(x)\log D(x)dx + \int_z p(z)\log(1 - D(G(z)))dz \\
=& \int_x p_{data}(x)\log D(x)dx + \int_x p_{model}(x)\log(1 - D(x))dx \\
=& \int_x \left[p_{data}(x)\log D(x) + p_{model}(x)\log(1 - D(x))\right]dx
\end{aligned}
\tag{37}
$$

Obviously, $\forall (a, b) \in \mathbb{R}^2 \backslash (0, 0)$, function $y = a\log x + b\log(1 - x)$ achieves its maximum in $[0, 1]$ at

$$
x = \frac{a}{a + b}
$$

And noted that $D(x)$ does not need to be defined outside of $Supp(p_{data}) \cup Supp(p_{model})$, where $Supp(\cdot)$ is the support of a distribution. Thus the optimal $D(x)$ is

$$
D(x) = \frac{p_{data}(x)}{p_{data}(x) + p_{model}(x)} \tag{38}
$$

$\square$

5 EM for mixture of multinomials

Problem 4

Given the corpus T and the corresponding vocabulary W, We denote by $T \in \mathbb{N}^{D\times W}$ the word occurrence matrix where the w-th word appears T_{dw} times in the d-th document. And each document d have a topic c_d in the topic-set with toal topic number of K. Words follows a multinomial distribution $Mult(\boldsymbol{\mu}_k)$, where $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2}, \cdots, \mu_{kW})^T$.

Thus

$$P(d|c_d = k) = \frac{n_d!}{\prod_{w=1}^{W} T_{dw}!} \prod_{w=1}^{W} \mu_{kw}^{T_{dw}} \tag{39}$$

where $n_d = \sum_{w=1}^{W} T_{dw}$.

Document d's distribution:

$$P(d) = \sum_{k=1}^{K} P(d|c_d = k)P(c_d = K) = \frac{n_d!}{\prod_{w=1}^{W} T_{dw}!} \sum_{k=1}^{K} \pi_k \prod_{w=1}^{W} \mu_{kw}^{T_{dw}} \tag{40}$$

Follwing EM algorithm, we can have:

E-step: Cauculate $\gamma_{dk} = P(c_d = k|d)$

$$\begin{aligned}\gamma_{dk} = P(c_d = k|d) &= \frac{P(d, c_d = k)}{P(d)} \\ &= \frac{\frac{n_d!}{\prod_{w=1}^{W} T_{dw}!}\pi_k \prod_{w=1}^{W} \mu_{kw}^{T_{dw}}}{\frac{n_d!}{\prod_{w=1}^{W} T_{dw}!} \sum_{j=1}^{K} \pi_j \prod_{w=1}^{W} \mu_{jw}^{T_{dw}}} \\ &= \frac{\pi_k \prod_{w=1}^{W} \mu_{kw}^{T_{dw}}}{\sum_{j=1}^{K} \pi_j \prod_{w=1}^{W} \mu_{jw}^{T_{dw}}}\end{aligned} \tag{41}$$

And γ_{dk} is fixed in the following **M-step**.

M-step: Maximum the Log Likelyhood's Expectation.

Let $\boldsymbol{D}$ be the set of documents, $\boldsymbol{C}$ be the set of topics.

$$\begin{aligned}L = \mathbb{E}_{\boldsymbol{C}}[\log P(\boldsymbol{D}, \boldsymbol{C})] &= \sum_{d=1}^{D} \mathbb{E}_{\boldsymbol{C}}[\log P(d, c_d)] \\ &= \sum_{d=1}^{D}\sum_{k=1}^{K} \log P(d, c_d = k)P(c_d = k|d) \\ &= \sum_{d=1}^{D}\sum_{k=1}^{K} \gamma_{dk} \left[\log \pi_k + \sum_{w=1}^{W} T_{dw} \log \mu_{kw}\right] \\ s.t. \quad & \sum_{k=1}^{K} \pi_k = 1 \\ & \sum_{w=1}^{W} \mu_{kw} = 1, k = 1, 2, \cdots, K\end{aligned} \tag{42}$$

By using Lagrange multiplier:

$$L(\boldsymbol{\pi}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{C}}[\log P(\boldsymbol{D}, \boldsymbol{C})] - \lambda_0 \left(\sum_{k=1}^{K} \pi_k - 1\right) - \sum_{k=1}^{K} \lambda_k \left(\sum_{w=1}^{W} \mu_{kw} - 1\right) \tag{43}$$

We have,

$$\begin{aligned}
\frac{\partial L}{\partial \pi_k} &= \frac{\sum_{d=1}^{D} \gamma_{dk}}{\pi_k} - \lambda_0 = 0 \\
\frac{\partial L}{\partial \mu_{kw}} &= \frac{\sum_{d=1}^{D} \gamma_{dk} T_{dw}}{\mu_{kw}} - \lambda_k = 0 \\
\sum_{k=1}^{K} \pi_k &= 1 \\
\sum_{w=1}^{W} \mu_{kw} &= 1, k = 1, 2, \cdots, K
\end{aligned} \tag{44}$$

so

$$\begin{aligned}
\pi_k &= \frac{\sum_{d=1}^{D} \gamma_{dk}}{\sum_{d=1}^{D} \sum_{k=1}^{K} \gamma_{dk}} \\
\mu_{kw} &= \frac{\sum_{d=1}^{D} \gamma_{dk} T_{dw}}{\sum_{d=1}^{D} \sum_{w=1}^{W} \gamma_{dk} T_{dw}}
\end{aligned} \tag{45}$$

□

In conclusion, we design EM for mixture of multinomials as below:
E-step: Cauculate γ_{dk} using the last iteration's model parameters π_k and μ_{kw}.

$$\gamma_{dk} = \frac{\pi_k \prod_{w=1}^{W} \mu_{kw}^{T_{dw}}}{\sum_{j=1}^{K} \pi_j \prod_{w=1}^{W} \mu_{jw}^{T_{dw}}} \tag{46}$$

M-step: Calculate and update model parameters π_k and μ_{kw}.

$$\begin{aligned}
\pi_k &= \frac{\sum_{d=1}^{D} \gamma_{dk}}{\sum_{d=1}^{D} \sum_{k=1}^{K} \gamma_{dk}} \\
\mu_{kw} &= \frac{\sum_{d=1}^{D} \gamma_{dk} T_{dw}}{\sum_{d=1}^{D} \sum_{w=1}^{W} \gamma_{dk} T_{dw}}
\end{aligned} \tag{47}$$

EM will repeat E-step and M-step until convergence. The iteration stop condition could be $||(\boldsymbol{\pi}, \boldsymbol{\mu})_{old} - (\boldsymbol{\pi}, \boldsymbol{\mu})_{new}||_2 < \epsilon$, where ϵ is very small number like 10^{-3}.

Problem 5

I implemented the EM algotithm using **Python**. The source file is `./code/em.py`. You may switch to directory `./code` and type `'python em.py --k K'` in your local terminal to run EM with `K` topics. Type `'python em.py --help'` to see more information of arguments.

I set the stop condition as $||\boldsymbol{\pi}_{old} - \boldsymbol{\pi}_{new}||_2 < 10^{-3}$.

Table 1 shows the iterations used until convergence and most-frequent words in each topic for $K = 10, 20, 30, 50$.

Table 1: iterations used until convergence and most-frequent words in each topic for $K = 10, 20, 30, 50$

K	#iter	topic t_k(**top5**-frequent words of t_k)[π_k] ($k \leq K$)
10	10	t_1(believe, point, really, going, read)[0.1544] t_2(drive, thanks, card, problem, using)[0.1472] t_3(available, file, information, program, data)[0.1245] t_4(government, year, years, law, really)[0.1191] t_5(game, team, year, problem, years)[0.0933] t_6(year, going, believe, point, years)[0.0879] t_7(image, space, data, years, using)[0.0804] t_8(window, windows, problem, using, server)[0.0801] t_9(jews, israel, turkey, game, problem)[0.0624] t_{10}(file, jpeg, image, windows, armenian)[0.0507]
20	9	t_1(image, data, thanks, software, available)[0.0817] t_2(really, doesnt, true, thanks, problem)[0.0633] t_3(thanks, windows, card, file, help)[0.0602] t_4(game, israel, point, going, better)[0.0600] t_5(going, didnt, drive, myers, problem)[0.0593] $\vdots$
30	7	t_1(believe, doesnt, point, really, going)[0.0544] t_2(thanks, windows, program, drive, really)[0.0539] t_3(drive, card, drives, disk, hard)[0.0525] t_4(windows, problem, drive, thanks, mhz)[0.0503] t_5(computer, windows, software, key, information)[0.0502] $\vdots$
50	7	t_1(drive, card, scsi, tape, thanks)[0.0392] t_2(windows, thanks, dos, file, software)[0.0359] t_3(problem, xfree, using, card, car)[0.0336] t_4(thanks, windows, best, really, offer)[0.0330] t_5(game, games, team, play, years)[0.0318] $\vdots$