CPEN455: Deep Learning
Homework 1

Mercury Mcindoe 85594505

January 19th 2025

1 Problem 1

1.1

Solution:

1.2

Solution:

Let's first consider the case before Dropout ($i.e., \mathbf{h}$). Since $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I)$, each entry x_i within $\mathbf{x}$ follows a normal distribution $\mathcal{N}(0,1)$ and each are iid. Let $\mathbf{z} = W\mathbf{x}$,

$$\mathbb{E}[\mathbf{z}] = \mathbb{E}[W\mathbf{x}] = W\mathbb{E}[\mathbf{x}] = 0$$

$$\mathrm{Var}(\mathbf{z}) = \mathrm{Var}(W\mathbf{x}) = W\mathrm{Var}(\mathbf{x})W^T = W \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) & \cdots & \mathrm{Cov}(x_1, x_N) \\ \mathrm{Cov}(x_2, x_1) & \mathrm{Var}(x_2) & \cdots & \mathrm{Cov}(x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_n, x_1) & \mathrm{Cov}(x_n, x_2) & \cdots & \mathrm{Var}(x_n, x_n) \end{bmatrix} W^T$$

Since we know that each x_i is iid which follows $\mathcal{N}(0,1)$, the variance-covariance matrix of $\mathbf{z}$ is then,

$$\mathrm{Var}(\mathbf{z}) = W \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} W^T = WW^T = I$$

Which shows that $\mathbf{z} = W\mathbf{x} \sim \mathcal{N}(0, I)$. Now considering that $\sigma(\mathbf{z}) = \max(\mathbf{z}, 0)$, each entry z_i would have the following expectations and variances,

$$\mathbb{E}[h_i] = \int_0^\infty \frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \left[-\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \right]_0^\infty = \frac{1}{\sqrt{2\pi}}$$

$$E[h_i^2] = \int_0^\infty \frac{z^2}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \frac{1}{2}$$

$$\mathrm{Var}(h_i) = \mathbb{E}[h_i^2] - (\mathbb{E}[h_i])^2 = \frac{1}{2} - \frac{1}{2\pi}$$

Putting all these together,

$$\therefore \mathrm{Var}(\mathbf{h}) = \left(\frac{1}{2} - \frac{1}{2\pi}\right) I_M$$

Now let's consider after Dropout, we know that $\tilde{\mathbf{h}} = \frac{\mathbf{m}}{1-p} \odot \mathbf{h}$, for an try $\tilde{h}_i$,

$$\begin{aligned}
\mathbb{E}[\tilde{h}_i] &= \mathbb{E}\left[\frac{m_i}{1-p} \cdot h_i\right] \\
&= \frac{1}{1-p}\mathbb{E}[m_i] \cdot \mathbb{E}[h_i] \\
&= \frac{1-p}{1-p}\mathbb{E}[h_i] \\
&= \frac{1}{\sqrt{2\pi}}.
\end{aligned}$$

The variance of $\tilde{h}_i$ is:

$$\begin{aligned}
\mathrm{Var}(\tilde{h}_i) &= \mathbb{E}[\tilde{h}_i^2] - \mathbb{E}[\tilde{h}_i]^2 \\
&= \mathbb{E}\left[\left(\frac{1}{1-p}\right)^2 \cdot m_i^2 \cdot h_i^2\right] - \left(\mathbb{E}\left[\frac{1}{1-p} \cdot m_i \cdot h_i\right]\right)^2 \\
&= \left(\frac{1}{1-p}\right)^2 \left(\mathbb{E}[m_i^2]\mathbb{E}[h_i^2] - \mathbb{E}[m_i]^2 \cdot \mathbb{E}[h_i]^2\right) \\
&= \frac{1}{1-p} \cdot \mathbb{E}[h_i^2] - \mathbb{E}[h_i]^2 \\
&= \frac{1}{1-p} \cdot \frac{1}{2} - \frac{1}{2\pi}.
\end{aligned}$$

Hence,

$$\therefore \mathrm{Var}(\tilde{\mathbf{h}}) = \left(\frac{1}{1-p} \cdot \frac{1}{2} - \frac{1}{2\pi}\right) I_M$$

1.3

Solution:

For one unit, the expectation that it is kept is $1-p$, then given M units the expectation would be $M \cdot (1-p)$ units kept. For each unit, we have a probability $1-p$ that it is kept, so if we compute the probabilty that k units are kept, $P(\text{kept} = k)$,

$$P(\text{kept} = k) = \binom{M}{k} \cdot (1-p)^k \cdot p^{M-k}$$

hence, a binomial distribution with probability $1-p$.

1.4

Solution:

First let $M(1-p) = \alpha$, in other words $p = 1 - \frac{\alpha}{M}$,

$$\lim_{M\to\infty} \binom{M}{k} \cdot (1-p)^k \cdot p^{M-k} = \lim_{M\to\infty} \frac{M(M-1)\cdot(M-k+1)}{k!}(1-p)^k(1-\frac{\alpha}{M})^{M-k}$$

$$= \lim_{M\to\infty} \frac{(M\cdot(1-p))\cdot((M-1)\cdot(1-p))\cdots((M-k+1)(1-p))}{k!}\cdot(1-\frac{\alpha}{M})^{M-k}$$

$$= \lim_{M\to\infty} \frac{(M\cdot(1-p))\cdot((M-1)\cdot(1-p))\cdots((M-k+1)(1-p))}{k!}\cdot(1-\frac{\alpha}{M})^{\frac{M}{\alpha}}\cdot(1-\frac{\alpha}{M})^{-k}$$

$$= \frac{\alpha^k}{k!}e^{-\alpha} = \frac{(M(1-p))^k}{k!}e^{-M(1-p)}$$

It becomes a Poisson distribution with parameter $M(1-p)$.

1.5

Solution:

Let's say that we want to keep x units and get the probability distribution. We then want to sum all the probabilities of keeping x units for all M. Thus we want,

$$P(x \text{ units kept}) = \sum_{M=x}^{\infty} P(x \text{ units kept} \cap M \text{ units})$$

Let's do the math!

$$\begin{aligned} P(x \text{ units kept} \cap M \text{ units}) &= \frac{\lambda^M e^{-\lambda}}{M!}\cdot\binom{M}{x}\cdot(1-p)^x\cdot p^{M-x} \\ &= \frac{\lambda^M e^{-\lambda}}{M!}\cdot\frac{M!}{(M-x)!\cdot x!}\cdot(1-p)^x\cdot p^{M-x} \\ &= \frac{e^{-\lambda}}{x!}\cdot(1-p)^x\cdot\frac{\lambda^M\cdot p^{M-x}}{(M-x)!}. \end{aligned}$$

Now, the probability of x units being kept is:

$$\begin{aligned} P(x \text{ units kept}) &= \sum_{M=x}^{\infty}\frac{e^{-\lambda}}{x!}\cdot(1-p)^x\cdot\frac{\lambda^M\cdot p^{M-x}}{(M-x)!} \\ &= \frac{e^{-\lambda}}{x!}\cdot(1-p)^x\cdot\sum_{M=x}^{\infty}\frac{\lambda^M\cdot p^{M-x}}{(M-x)!}. \end{aligned}$$

Let $M' = M - x$. Then:

$$\begin{aligned} P(x \text{ units kept}) &= \frac{e^{-\lambda}}{x!}\cdot(1-p)^x\cdot\sum_{M'=0}^{\infty}\frac{\lambda^{M'+x}\cdot p^{M'}}{M'!} \\ &= \frac{e^{-\lambda}}{x!}\cdot(1-p)^x\cdot\lambda^x\cdot\sum_{M'=0}^{\infty}\frac{(\lambda p)^{M'}}{M'!}. \end{aligned}$$

Since $\sum_{M'=0}^{\infty} \frac{(\lambda p)^{M'}}{M'!} = e^{\lambda p}$, we get:

$$\begin{aligned} P(x \text{ units kept}) &= \frac{e^{-\lambda}}{x!} \cdot (1-p)^x \cdot \lambda^x \cdot e^{\lambda p} \\ &= \frac{e^{-\lambda(1-p)} \cdot \{\lambda(1-p)\}^x}{x!}. \end{aligned}$$

Therefore, the number of kept units follows a Poisson distribution with parameter $\lambda(1-p)$.

2 Problem 2

2.1

Solution:

We need the hyperparameter ϵ to avoid division by 0 during training.

2.2

Solution:

$$\begin{aligned} \mathbb{E}[\hat{Y}[i,j]] &= \mathbb{E}\left[\gamma[j] \cdot \frac{Y[i,j] - \mathbf{m}[j]}{\sqrt{\mathbf{v}[j]}} + \boldsymbol{\beta}[j]\right] \\ &= \frac{\gamma[j]}{\mathbf{v}[j]} \cdot (\mathbb{E}\left[Y[i,j]\right] - \mathbb{E}[\mathbf{m}[j]]) + \mathbb{E}[\boldsymbol{\beta}[j]] \\ &= \frac{\gamma[j]}{\mathbf{v}[j]} \cdot (\mathbf{m}[j] - \mathbf{m}[j]) + \boldsymbol{\beta}[j] \\ &= \boldsymbol{\beta}[j] \end{aligned}$$

$$\begin{aligned} Var(\hat{Y}[i,j]) &= \mathbb{E}[\hat{Y}[i,j]^2] - \mathbb{E}[\hat{Y}[i,j]]^2 = \mathbb{E}[\hat{Y}[i,j]^2] - \boldsymbol{\beta}[j]^2 \\ &= \mathbb{E}\left[\frac{\gamma[j]^2}{\mathbf{v}[j]} \cdot (Y[i,j] - \mathbf{m}[j])^2 + 2 \cdot \frac{\gamma[j]\boldsymbol{\beta}[j]}{\sqrt{\mathbf{v}[j]}} \cdot (E\left[Y[i,j] - \mathbf{m}[j]\right]) + \boldsymbol{\beta}[j]^2\right] - \boldsymbol{\beta}[j]^2 \\ &= \frac{\gamma[j]^2}{\mathbf{v}[j]} \cdot \mathbb{E}\left[(Y(i,j) - \mathbf{m}[j])^2\right] + 0 + 0 \\ &= \frac{\gamma[j]^2}{\mathbf{v}[j]} \cdot \mathbb{E}\left[Y[i,j]^2 - 2Y[i,j]\mathbf{m}[j] + \mathbf{m}[j]^2\right] \\ &= \frac{\gamma[j]^2}{\mathbf{v}[j]} \cdot \left(\mathbb{E}\left[Y[i,j]^2\right] - 2\mathbf{m}[j]^2 + \mathbf{m}[j]^2\right) \\ &= \frac{\gamma[j]^2}{\mathbf{v}[j]} \cdot \left(\mathbf{v}[j] - \mathbf{m}[j]^2 + \mathbf{m}[j]^2\right) \\ &= \gamma[j]^2 \end{aligned}$$

Thus,

$$\mathbb{E}\left[\hat{Y}[i,j]\right] = \boldsymbol{\beta}[j] \quad Var(\hat{Y}[i,j]) = \gamma[j]^2$$

2.3

Solution:

Firstly, let's denote the derivative of the ReLU activation function, we denote it as f where f denotes the step function from 0 to 1 where the scalar input equals zero. Also, we will denote the indexes with subscripts, for instance, $Y[i,j] = Y_{ij}$ in the following explanations.

Problem 1: $\frac{\partial \ell}{\partial \gamma}$

$$\frac{\partial \ell}{\partial \gamma} = \begin{bmatrix} \frac{\partial \ell}{\partial \gamma_1} \\ \frac{\partial \ell}{\partial \gamma_2} \\ \vdots \\ \frac{\partial \ell}{\partial \gamma_M} \end{bmatrix}$$

Now let's compute $\frac{\partial \ell}{\partial \gamma_j}$ for $j = 1, \ldots, M$.

$$\frac{\partial \ell}{\partial \gamma_j} = \sum_{k=1}^{B} \sum_{l=1}^{M} \frac{\partial \ell}{\partial H_{kl}} \frac{\partial H_{kl}}{\partial \hat{Y}_{kl}} \frac{\partial \hat{Y}_{kl}}{\partial \gamma_j}$$

$$= \sum_{k=1}^{B} \sum_{l=1}^{M} \frac{\partial \ell}{\partial H_{kl}} f(\hat{Y}_{kl}) \frac{\partial \hat{Y}_{kl}}{\partial \gamma_j}$$

Now let's compute $\frac{\partial \hat{Y}_{kl}}{\partial \gamma j}$,

$$\begin{aligned} \frac{\partial \hat{Y}_{kl}}{\partial \gamma_j} &= \frac{\partial}{\partial \gamma_j} \left(\gamma_l \cdot \frac{Y_{kl} - \mathbf{m}_l}{\sqrt{\mathbf{v}_l + \epsilon}} + \boldsymbol{\beta}_j \right) \\ &= \delta_{lj} \cdot \frac{Y_{kl} - \mathbf{m}_l}{\sqrt{\mathbf{v}_l + \epsilon}} \end{aligned}$$

In summary,

$$\begin{aligned} \frac{\partial \ell}{\partial \gamma_j} &= \sum_{k=1}^{B} \sum_{l=1}^{M} \frac{\partial \ell}{\partial H_{kl}} f(\hat{Y}_{kl}) \frac{\partial \hat{Y}_{kl}}{\partial \gamma_j} \\ &= \sum_{k=1}^{B} \frac{\partial \ell}{\partial H_{kj}} f(\hat{Y}_{kj}) \frac{Y_{kj} - \mathbf{m}_j}{\sqrt{\mathbf{v}_j + \epsilon}} \end{aligned}$$

Problem 2: $\frac{\partial l}{\partial \beta}$

$$\frac{\partial l}{\partial \beta} = \begin{bmatrix} \frac{\partial l}{\partial \beta_1} \\ \frac{\partial l}{\partial \beta_2} \\ \vdots \\ \frac{\partial l}{\partial \beta_M} \end{bmatrix}$$

Similarly, let's compute $\frac{\partial l}{\partial \beta_j}$ for $j = 1, \cdots, M$.

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{k=1}^{B}\sum_{l=1}^{M} \frac{\partial \ell}{\partial H_{kl}} \frac{\partial H_{kl}}{\partial \hat{Y}_{kl}} \frac{\partial \hat{Y}_{kl}}{\partial \beta_j}$$

$$= \sum_{k=1}^{B}\sum_{l=1}^{M} \frac{\partial \ell}{\partial H_{kl}} f(\hat{Y}_{kl}) \frac{\partial \hat{Y}_{kl}}{\partial \beta_j}$$

Now we compute $\frac{\partial \hat{Y}_{kl}}{\partial \beta_j}$,

$$\begin{aligned}\frac{\partial \hat{Y}_{kl}}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\left(\gamma_l \cdot \frac{Y_{kl} - \mathbf{m}_l}{\sqrt{\mathbf{v}_l + \epsilon}} + \beta_l\right) \\ &= \delta_{lj}\end{aligned}$$

In summary,

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \sum_{k=1}^{B}\sum_{l=1}^{M} \frac{\partial \ell}{\partial H_{kl}} f(\hat{Y}_{kl}) \frac{\partial \hat{Y}_{kl}}{\partial \beta_j} \\ &= \sum_{k=1}^{B} \frac{\partial \ell}{\partial H_{kj}} f(\hat{Y}_{kj})\end{aligned}$$

Problem 3: $\frac{\partial \ell}{\partial \mathbf{m}}$

$$\frac{\partial \ell}{\partial \mathbf{m}} = \begin{bmatrix} \frac{\partial \ell}{\partial \mathbf{m}_1} \\ \frac{\partial \ell}{\partial \mathbf{m}_2} \\ \vdots \\ \frac{\partial \ell}{\partial \mathbf{m}_M} \end{bmatrix}$$

Let's find $\frac{\partial \ell}{\partial \mathbf{m}_j}$ for $j = 1, \ldots, M$.

$$\frac{\partial \ell}{\partial \mathbf{m}_j} = \sum_{k=1}^{B}\sum_{l=1}^{M} \frac{\partial \ell}{\partial H_{kl}} \frac{\partial H_{kl}}{\partial \hat{Y}_{kl}} \frac{\partial \hat{Y}_{kl}}{\partial \mathbf{m}_j}$$

$$= \sum_{k=1}^{B}\sum_{l=1}^{M} \frac{\partial \ell}{\partial H_{kl}} f(\hat{Y}_{kl}) \frac{\partial \hat{Y}_{kl}}{\partial \mathbf{m}_j}$$

If we compute $\frac{\partial \hat{Y}_{kl}}{\partial \mathbf{m}_j}$,

$$\begin{aligned}\frac{\partial \hat{Y}_{kl}}{\partial \mathbf{m}_j} &= \frac{\partial}{\partial \mathbf{m}_j}\left(\gamma_l \cdot \frac{Y_{kl} - \mathbf{m}_l}{\sqrt{\mathbf{v}_l + \epsilon}} + \beta_l\right) \\ &= -\delta_{lj} \cdot \gamma_l \cdot \frac{1}{\sqrt{\mathbf{v}_l + \epsilon}} + \gamma_l \cdot (Y_{kl} - \mathbf{m}_l) \cdot (-\frac{1}{2}) \cdot (\mathbf{v}_l + \epsilon)^{-\frac{3}{2}} \cdot \frac{\partial \mathbf{v}_l}{\partial \mathbf{m}_j}\end{aligned}$$

Let's investigate what $\frac{\partial \mathbf{v}_l}{\partial \mathbf{m}_j}$ is,

$$
\begin{aligned}
\frac{\partial \mathbf{v}_l}{\partial \mathbf{m}_j} &= \frac{\partial}{\partial \mathbf{m}_j}\left(\frac{1}{B}\sum_{i=1}^{B}(Y_{il}-\mathbf{m}_l)^2\right) \\
&= \frac{2}{B}\sum_{i=1}^{B}(Y_{il}-\mathbf{m}_l)\cdot\frac{\partial \mathbf{m}_l}{\partial \mathbf{m}_j} \\
&= \delta_{lj}\cdot\frac{2}{B}\sum_{i=1}^{B}(Y_{il}-\mathbf{m}_l) \\
&= 0
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\frac{\partial \ell}{\partial \mathbf{m}_j} &= \sum_{k=1}^{B}\sum_{l=1}^{M}\frac{\partial \ell}{\partial H_{kl}} f(\hat{Y}_{kl})\frac{\partial \hat{Y}_{kl}}{\partial \mathbf{m}_j} \\
&= \sum_{k=1}^{B}\sum_{l=1}^{M}\frac{\partial \ell}{\partial H_{kl}} f(\hat{Y}_{kl})\left(-\delta_{lj}\cdot\gamma_l\cdot\frac{1}{\sqrt{\mathbf{v}_l+\epsilon}}\right) \\
&= \sum_{k=1}^{B}\frac{\partial \ell}{\partial H_{kj}} f(\hat{Y}_{kj})\left(-\gamma_j\cdot\frac{1}{\sqrt{\mathbf{v}_j+\epsilon}}\right)
\end{aligned}
$$

Problem 4: $\frac{\partial \ell}{\partial \mathbf{v}}$

$$
\frac{\partial \ell}{\partial \mathbf{v}} = \begin{bmatrix} \frac{\partial \ell}{\partial \mathbf{v}_1} \\ \frac{\partial \ell}{\partial \mathbf{v}_2} \\ \vdots \\ \frac{\partial \ell}{\partial \mathbf{v}_M} \end{bmatrix}
$$

As usual, let's compute $\frac{\partial \ell}{\partial \mathbf{v}_j}$ for $j = 1, \ldots, M$.

$$
\begin{aligned}
\frac{\partial \ell}{\partial \mathbf{v}_j} &= \sum_{k=1}^{B}\sum_{l=1}^{M}\frac{\partial \ell}{\partial H_{kl}}\frac{\partial H_{kl}}{\partial \hat{Y}_{kl}}\frac{\partial \hat{Y}_{kl}}{\partial \mathbf{v}_j} \\
&= \sum_{k=1}^{B}\sum_{l=1}^{M}\frac{\partial \ell}{\partial H_{kl}} f(\hat{Y}_{kl})\frac{\partial \hat{Y}_{kl}}{\partial \mathbf{v}_j}
\end{aligned}
$$

Let's compute $\frac{\partial \hat{Y}_{kl}}{\partial \mathbf{v}_j}$,

$$
\begin{aligned}
\frac{\partial \hat{Y}_{kl}}{\partial \mathbf{v}_j} &= \frac{\partial}{\partial \mathbf{v}_j}\left(\gamma_l\cdot\frac{Y_{kl}-\mathbf{m}_l}{\sqrt{\mathbf{v}_l+\epsilon}}+\beta_l\right) \\
&= \gamma_l\cdot(Y_{kl}-\mathbf{m}_l)\cdot(-\frac{1}{2})\cdot(\mathbf{v}_l+\epsilon)^{-\frac{3}{2}}\cdot\frac{\partial \mathbf{v}_l}{\partial \mathbf{v}_j} \\
&= \delta_{lj}\cdot\gamma_l\cdot(Y_{kl}-\mathbf{m}_l)\cdot(-\frac{1}{2})\cdot(\mathbf{v}_l+\epsilon)^{-\frac{3}{2}}
\end{aligned}
$$

Therefore,

$$\begin{aligned}
\frac{\partial \ell}{\partial \mathbf{v}_j} &= \sum_{k=1}^{B}\sum_{l=1}^{M} \frac{\partial \ell}{\partial H_{kl}} f(\hat{Y}_{kl}) \frac{\partial \hat{Y}_{kl}}{\partial \mathbf{v}_j} \\
&= \sum_{k=1}^{B}\sum_{l=1}^{M} \frac{\partial \ell}{\partial H_{kl}} f(\hat{Y}_{kl}) \left(\delta_{lj} \cdot \gamma_l \cdot (Y_{kl} - \mathbf{m}_l) \cdot (-\frac{1}{2}) \cdot (\mathbf{v}_l + \epsilon)^{-\frac{3}{2}}\right) \\
&= \sum_{k=1}^{B} \frac{\partial \ell}{\partial H_{kj}} f(\hat{Y}_{kj}) \left(-\gamma_j \cdot \frac{1}{2} \cdot (Y_{kj} - \mathbf{m}_j) \cdot (\mathbf{v}_j + \epsilon)^{-\frac{3}{2}}\right)
\end{aligned}$$

Problem 5: $\frac{\partial \ell}{\partial Y}$

$$\frac{\partial \ell}{\partial Y} = \begin{bmatrix} \frac{\partial \ell}{\partial Y_{11}} & \cdots & \frac{\partial \ell}{\partial Y_{1M}} \\ \vdots & \frac{\partial \ell}{\partial Y_{st}} & \vdots \\ \frac{\partial \ell}{\partial Y_{B1}} & \cdots & \frac{\partial \ell}{\partial Y_{BM}} \end{bmatrix}$$

Now, we compute $\frac{\partial \ell}{\partial Y_{st}}$ for $s = 1, \ldots, B$ and $t = 1, \ldots, M$.

$$\frac{\partial \ell}{\partial Y_{st}} = \sum_{k=1}^{B}\sum_{l=1}^{M} \frac{\partial \ell}{\partial H_{kl}} f(\hat{Y}_{kl}) \frac{\partial \hat{Y}_{kl}}{\partial Y_{st}}$$

Let's compute $\frac{\partial \hat{Y}_{kl}}{\partial Y_{st}}$,

$$\begin{aligned}
\frac{\partial \hat{Y}_{kl}}{\partial Y_{st}} &= \frac{\partial}{\partial Y_{st}} \left(\gamma_l \cdot \frac{Y_{kl} - \mathbf{m}_l}{\sqrt{\mathbf{v}_l + \epsilon}} + \beta_l\right) \\
&= \gamma_l \cdot (\frac{\partial Y_{kl}}{\partial Y_{st}} - \frac{\partial \mathbf{m}_l}{\partial Y_{st}}) \cdot \frac{1}{\sqrt{\mathbf{v}_l + \epsilon}} - \frac{1}{2} \cdot \gamma_l \cdot (Y_{kl} - \mathbf{m}_l) \cdot (\mathbf{v}_l + \epsilon)^{-\frac{3}{2}} \cdot \frac{\partial \mathbf{v}_l}{\partial Y_{st}}
\end{aligned}$$

We know have to find out $\frac{\partial Y_{kl}}{\partial Y_{st}}, \frac{\partial \mathbf{m}_l}{\partial Y_{st}}, \frac{\partial \mathbf{v}_l}{\partial Y_{st}}$,

$$\frac{\partial Y_{kl}}{\partial Y_{st}} = \delta_{ks} \cdot \delta_{lt}$$

$$\begin{aligned}
\frac{\partial \mathbf{m}_l}{\partial Y_{st}} &= \frac{\partial}{\partial Y_{st}} \left(\frac{1}{B}\sum_{i=1}^{B} Y_{il}\right) \\
&= \frac{1}{B} \cdot \delta_{lt}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathbf{v}_l}{\partial Y_{st}} &= \frac{\partial}{\partial Y_{st}}\left(\frac{1}{B}\sum_{i=1}^{B}(Y_{il}-\mathbf{m}_l)^2\right) \\
&= \frac{2}{B}\sum_{i=1}^{B}(Y_{il}-\mathbf{m}_l)\cdot\left(\frac{\partial Y_{il}}{\partial Y_{st}}-\frac{\partial \mathbf{m}_l}{\partial Y_{st}}\right) \\
&= \frac{2}{B}\sum_{i=1}^{B}(Y_{il}-\mathbf{m}_l)\cdot(\delta_{is}\cdot\delta_{lt}-\frac{1}{B}\cdot\delta_{lt}) \\
&= \frac{2}{B}\cdot\delta_{lt}\cdot(Y_{sl}-\mathbf{m}_l)-\frac{2}{B^2}\cdot\delta_{lt}\cdot\sum_{i=1}^{B}(Y_{il}-\mathbf{m}_l) \\
&= \frac{2}{B}\cdot\delta_{lt}\cdot(Y_{sl}-\mathbf{m}_l)
\end{aligned}$$

Putting it all together,

$$\frac{\partial \hat{Y}_{kl}}{\partial Y_{st}} = \gamma_l\cdot(\delta_{ks}\cdot\delta_{lt}-\frac{1}{B}\cdot\delta_{lt})\cdot\frac{1}{\sqrt{\mathbf{v}_l+\epsilon}}-\frac{1}{2}\cdot\gamma_l\cdot(Y_{kl}-\mathbf{m}_l)\cdot(\mathbf{v}_l+\epsilon)^{-\frac{3}{2}}\cdot\frac{2}{B}\cdot\delta_{lt}\cdot(Y_{sl}-\mathbf{m}_l)$$

And,

$$\begin{aligned}
\frac{\partial \ell}{\partial Y_{st}} &= \sum_{k=1}^{B}\sum_{l=1}^{M}\frac{\partial \ell}{\partial H_{kl}}f(\hat{Y}_{kl})\frac{\partial \hat{Y}_{kl}}{\partial Y_{st}} \\
&= \sum_{k=1}^{B}\sum_{l=1}^{M}\frac{\partial \ell}{\partial H_{kl}}f(\hat{Y}_{kl})\left\{\gamma_l\cdot(\delta_{ks}\cdot\delta_{lt}-\frac{1}{B}\cdot\delta_{lt})\cdot\frac{1}{\sqrt{\mathbf{v}_l+\epsilon}}-\frac{1}{2}\cdot\gamma_l\cdot(Y_{kl}-\mathbf{m}_l)\cdot(\mathbf{v}_l+\epsilon)^{-\frac{3}{2}}\cdot\frac{2}{B}\cdot\delta_{lt}\cdot(Y_{sl}-\mathbf{m}_l)\right\} \\
&= \sum_{k=1}^{B}\sum_{l=1}^{M}\frac{\partial \ell}{\partial H_{kl}}f(\hat{Y}_{kl})\left\{\gamma_l\cdot(\delta_{ks}\cdot\delta_{lt}-\frac{1}{B}\cdot\delta_{lt})\cdot(\mathbf{v}_l+\epsilon)^{-\frac{1}{2}}-\frac{1}{B}\cdot\gamma_l\cdot(Y_{kl}-\mathbf{m}_l)\cdot(\mathbf{v}_l+\epsilon)^{-\frac{3}{2}}\cdot\delta_{lt}\cdot(Y_{sl}-\mathbf{m}_l)\right\} \\
&= \frac{\partial \ell}{\partial H_{st}}f(\hat{Y}_{st})\cdot\gamma_t\cdot(\mathbf{v}_t+\epsilon)^{-\frac{1}{2}}-\frac{1}{B}\sum_{k=1}^{B}\frac{\partial \ell}{\partial H_{kt}}f(\hat{Y}_{kt})\cdot\gamma_t\cdot\left\{(\mathbf{v}_t+\epsilon)^{-\frac{1}{2}}-(Y_{kt}-\mathbf{m}_t)\cdot(\mathbf{v}_t+\epsilon)^{-\frac{3}{2}}\cdot(Y_{st}-\mathbf{m}_t)\right\}
\end{aligned}$$

3 Problem 3

3.1

Solution:

From class, we learned that,

$$\begin{aligned}
\frac{\partial \ell}{\partial \mathbf{h}_L} &= \left(\frac{\partial \mathbf{y}}{\partial \mathbf{h}_L}\right)^T\frac{\partial \ell}{\partial \mathbf{y}} \\
\frac{\partial \ell}{\partial \mathbf{h}_{L-1}} &= \left(\frac{\partial \mathbf{h}_L}{\partial \mathbf{h}_{L-1}}\right)^T\frac{\partial \ell}{\partial \mathbf{h}_L}
\end{aligned}$$

Then for $1 \le i \le L$, we can say that,

$$\frac{\partial \ell}{\partial \mathbf{h}_i} = \left(\frac{\partial \mathbf{h}_{i+1}}{\partial \mathbf{h}_i}\right)^T \left(\frac{\partial \mathbf{h}_{i+2}}{\partial \mathbf{h}_{i+1}}\right)^T \cdots \left(\frac{\partial \mathbf{y}}{\partial \mathbf{h}_L}\right)^T \frac{\partial \ell}{\partial \mathbf{y}}$$

Let's first compute $\frac{\partial \ell}{\partial \mathbf{y}}$,

$$\frac{\partial \ell}{\partial \mathbf{y}} = \begin{bmatrix} \frac{\partial \ell}{\partial \mathbf{y}_1} \\ \vdots \\ \frac{\partial \ell}{\partial \mathbf{y}_{D_L}} \end{bmatrix} \in \mathbb{R}^{D_L \times 1}$$

Then, for $\frac{\partial \ell}{\partial \mathbf{y}_j}$, for $j = 1, \ldots, D_L$.

$$\begin{aligned} \frac{\partial \ell}{\partial \mathbf{y}_j} &= \frac{\partial}{\partial \mathbf{y}_j}\left(-\sum_{k=1}^{K} \bar{\mathbf{y}}[k] \cdot \log(\mathbf{y}[k])\right) \\ &= -\sum_{k=1}^{K} \bar{\mathbf{y}}[k] \cdot \frac{\frac{\partial \mathbf{y}[k]}{\partial \mathbf{y}[j]}}{\mathbf{y}[k]} \\ &= -\frac{\bar{\mathbf{y}}[j]}{\mathbf{y}[j]} \end{aligned}$$

Now let's find $\frac{\partial \mathbf{h}_{k+1}}{\partial \mathbf{h}_k}$ for $k = i, \ldots, L-1$. We are aware that $\mathbf{h}_i = \sigma\left(W_i \mathbf{h}_{i-1} + \mathbf{b}_i\right)$. If we let $z_i = W_i \mathbf{h}_{i-1} + \mathbf{b}_i$, then,

$$\begin{aligned} \frac{\partial \mathbf{h}_{k+1}}{\partial \mathbf{h}_k} &= \frac{\partial \mathbf{h}_{k+1}}{\partial z_{k+1}} \cdot \frac{\partial z_{k+1}}{\partial \mathbf{h}_k} \\ &= \mathrm{diag}(\sigma'(z_{k+1})) \cdot W_{k+1} \in \mathbb{R}^{D_{k+1} \times D_k} \end{aligned}$$

Now, let's get $\frac{\partial \mathbf{y}}{\partial \mathbf{h}_L}$,

$$\frac{\partial \mathbf{y}}{\partial \mathbf{h}_L} = \begin{bmatrix} \frac{\partial \mathbf{y}[1]}{\partial \mathbf{h}_L[1]} & \cdots & \frac{\partial \mathbf{y}[1]}{\partial \mathbf{h}_L[D_L]} \\ \vdots & & \vdots \\ \frac{\partial \mathbf{y}[D_L]}{\partial \mathbf{h}_L[1]} & \cdots & \frac{\partial \mathbf{y}[D_L]}{\partial \mathbf{h}_L[D_L]} \end{bmatrix} \in \mathbb{R}^{D_L \times D_L}$$

So let's try computing $\frac{\partial \mathbf{y}_i}{\partial \mathbf{h}_L[k]}$,

$$\begin{aligned} \frac{\partial \mathbf{y}_i}{\partial \mathbf{h}_L[k]} &= \frac{\partial}{\partial \mathbf{h}_L[k]}\left(\frac{\exp(\mathbf{h}_L[i])}{\sum_j \exp(\mathbf{h}_L[j])}\right) \\ &= \frac{\partial \mathbf{h}_L[i]}{\partial \mathbf{h}_L[k]} \cdot \left(\frac{\exp(\mathbf{h}_L[i])}{\sum_j \exp(\mathbf{h}_L[j])}\right) - \exp(\mathbf{h}_L[i]) \cdot \frac{\sum_j \frac{\partial \mathbf{h}_L[j]}{\partial \mathbf{h}_L[k]} \cdot \exp(\mathbf{h}_L[j])}{\left(\sum_j \exp(\mathbf{h}_L[j])\right)^2} \\ &= \delta_{ki} \cdot \mathbf{y}[i] - \exp(\mathbf{h}_L[i]) \cdot \frac{\exp(\mathbf{h}_L[k])}{\left(\sum_j \exp(\mathbf{h}_L[j])\right)^2} \\ &= \delta_{ki} \cdot \mathbf{y}[i] - \mathbf{y}[i] \cdot \mathbf{y}[k] \end{aligned}$$

For simplicity, let's pre-compute $\left(\frac{\partial \mathbf{y}}{\partial \mathbf{h}_L}\right)^T \frac{\partial \ell}{\partial \mathbf{y}}$,

$$\left(\frac{\partial \mathbf{y}}{\partial \mathbf{h}_L}\right)^T \frac{\partial \ell}{\partial \mathbf{y}} = \begin{bmatrix} \mathbf{y}[1]-\mathbf{y}[1]^2 & -\mathbf{y}[2]\mathbf{y}[1] & \cdots & -\mathbf{y}[D_L]\mathbf{y}[1] \\ -\mathbf{y}[1]\mathbf{y}[2] & \mathbf{y}[2]-\mathbf{y}[2]^2 & \cdots & -\mathbf{y}[D_L]\mathbf{y}[2] \\ \vdots & \vdots & \cdots & \vdots \\ -\mathbf{y}[1]\mathbf{y}[D_L] & -\mathbf{y}[2]\mathbf{y}[D_L] & \cdots & \mathbf{y}[D_L]-\mathbf{y}[D_L]^2 \end{bmatrix} \begin{bmatrix} -\frac{\bar{\mathbf{y}}[1]}{\mathbf{y}[1]} \\ \vdots \\ -\frac{\bar{\mathbf{y}}[D_L]}{\mathbf{y}[D_L]} \end{bmatrix}$$
$$= \begin{bmatrix} -\bar{\mathbf{y}}[1] + \mathbf{y}[1]\left(\sum_j \bar{\mathbf{y}}[j]\right) \\ -\bar{\mathbf{y}}[2] + \mathbf{y}[2]\left(\sum_j \bar{\mathbf{y}}[j]\right) \\ \vdots \\ -\bar{\mathbf{y}}[D_L] + \mathbf{y}[D_L]\left(\sum_j \bar{\mathbf{y}}[j]\right) \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{y}[1]-\bar{\mathbf{y}}[1] \\ \mathbf{y}[2]-\bar{\mathbf{y}}[2] \\ \vdots \\ \mathbf{y}[D_L]-\bar{\mathbf{y}}[D_L] \end{bmatrix}$$

Thus,

$$\frac{\partial \ell}{\partial \mathbf{h}_i} = \left(\frac{\partial \mathbf{h}_{i+1}}{\partial \mathbf{h}_i}\right)^T \left(\frac{\partial \mathbf{h}_{i+2}}{\partial \mathbf{h}_{i+1}}\right)^T \cdots \left(\frac{\partial \mathbf{y}}{\partial \mathbf{h}_L}\right)^T \frac{\partial \ell}{\partial \mathbf{y}}$$
$$= (\mathrm{diag}(\sigma'(z_i)) \cdot W_i)^T (\mathrm{diag}(\sigma'(z_{i+1})) \cdot W_{i+1})^T \cdots (\mathrm{diag}(\sigma'(z_{L-1})) \cdot W_{L-1})^T \left(\begin{bmatrix} \mathbf{y}[1]-\bar{\mathbf{y}}[1] \\ \mathbf{y}[2]-\bar{\mathbf{y}}[2] \\ \vdots \\ \mathbf{y}[D_L]-\bar{\mathbf{y}}[D_L] \end{bmatrix}\right) \in \mathbb{R}^{D_i \times 1}$$

3.2

Solution:

Let's start with $\frac{\partial \ell}{\partial W_i} \in \mathbb{R}^{D_i \times D_{i-1}}$,

$$\frac{\partial \ell}{\partial W_i} = \begin{bmatrix} \frac{\partial \ell}{\partial W_i[1,1]} & \frac{\partial \ell}{\partial W_i[2,1]} & \cdots & \frac{\partial \ell}{\partial W_i[D_{i-1},1]} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial \ell}{\partial W_i[1,D_i]} & \cdots & \cdots & \frac{\partial \ell}{\partial W_i[D_{i-1},D_i]} \end{bmatrix}$$

Using the same setup as the prior question, $\mathbf{h}_i = \sigma(W_i \mathbf{h}_{i-1} + \mathbf{b}_i), z_i = W_i \mathbf{h}_{i-1} + \mathbf{b}_i$,

$$\frac{\partial \ell}{\partial W_i[x,y]} = \frac{\partial \ell}{\partial z_i[x]} \cdot \frac{\partial z_i[x]}{\partial W_i[x,y]}$$
$$= \frac{\partial \ell}{\partial z_i[x]} \cdot \mathbf{h}_{i-1}[y]$$

Given this, if we were to compute $\frac{\partial \ell}{\partial z_i}$,

$$\begin{aligned}\frac{\partial \ell}{\partial z_i} &= \left(\frac{\partial \mathbf{h}_i}{\partial z_i}\right)^T \frac{\partial \ell}{\partial \mathbf{h}_i} \\ &= \sigma'(z_i) \odot \frac{\partial \ell}{\partial \mathbf{h}_i}\end{aligned}$$

These combined will give us,

$$\frac{\partial \ell}{\partial W_i} = \left(\sigma'(z_i) \odot \frac{\partial \ell}{\partial \mathbf{h}_i}\right) \mathbf{h}_{i-1}^T \in \mathbb{R}^{D_i \times D_{i-1}}$$

where $\frac{\partial \ell}{\partial \mathbf{h}_i} \in \mathbb{R}^{D_i \times 1}$.
Now, let's compute $\frac{\partial \ell}{\partial \mathbf{b}_i}$,

$$\begin{aligned}\frac{\partial \ell}{\partial \mathbf{b}_i} &= \left(\frac{\partial z_i}{\partial \mathbf{b}_i}\right)^T \left(\frac{\partial \mathbf{h}_i}{\partial z_i}\right)^T \frac{\partial \ell}{\partial \mathbf{h}_i} \\ &= I\left(\sigma'(z_i) \odot \frac{\partial \ell}{\partial \mathbf{h}_i}\right) \\ &= \sigma'(z_i) \odot \frac{\partial \ell}{\partial \mathbf{h}_i} \in \mathbb{R}^{D_i \times 1}\end{aligned}$$