BME/ECE 695 Deep Learning
Midterm II Solution
April 23, Spring 2020

**Q1.**
2 Points

**Rules:** I understand that this is an open book exam that shall be done within the allotted time of 120 minutes. I can use my notes, and web resources. However, I will not communicate with any other person other than the official exam proctors during the exam, and I will not seek or accept help from any other persons other than the official proctors.

*Upload a scan of your signature here:*

**Name:** (4 pt) _____

**Q2 Back Propagation**
32 Points
Consider the inference function

$$f_\theta(y)$$

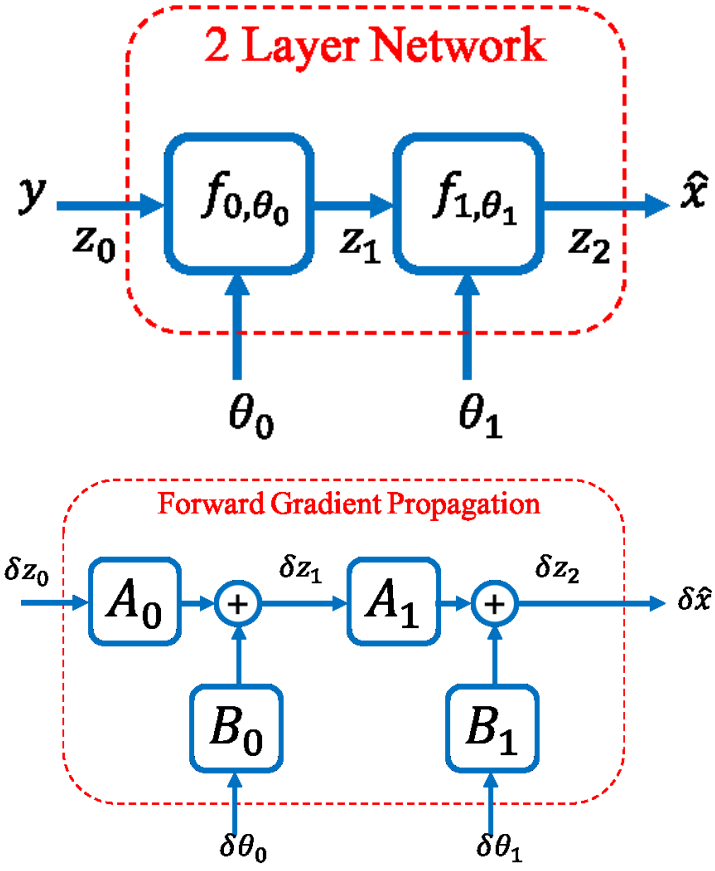
where

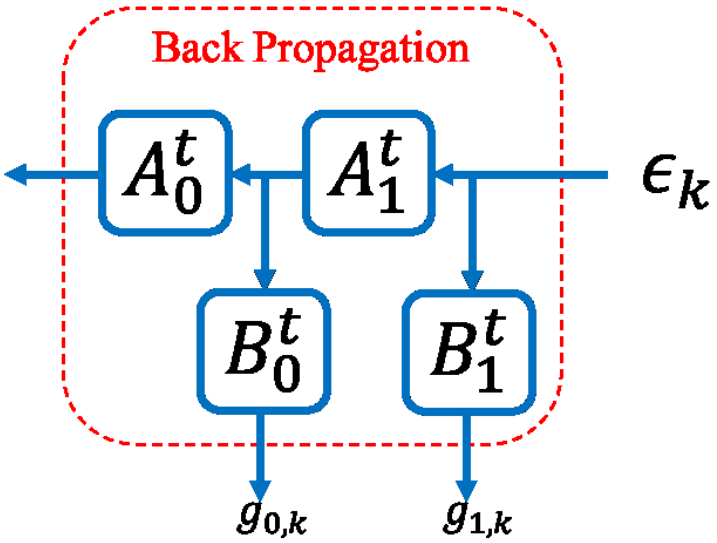
$$f_\theta(y) = f_{1,\theta_1}\left(f_{0,\theta_0}\left(y\right)\right)$$

and the associated loss function is given by

$$L(\theta) = \frac{1}{K}\sum_{k=0}^{K-1}\|x_k - f_\theta(y_k)\|^2$$

where $[x_k, y_k]$ for $k = 0, \cdots K-1$ are training pairs, and $\theta = [\theta_0, \theta_1]$ is the associated full parameter vector.
The three figures below illustrate: a) a 2 Layer Network; b) the Forward Gradient Propagation network; and c) the Back Propagation network.






**Q2.1**
8 Points For the forward gradient propagation with the training pair $[x_k, y_k]$, give expressions for $A_0$, $B_0$, $A_1$, and $B_1$ in terms of the functions $f_{0,\theta_0}(z_0)$ and $f_{1,\theta_1}(z_1)$.
—files—

**Q2.2**
8 Points
For the forward gradient propagation with the training pair $[x_k, y_k]$, give an expression for $\delta\hat{x}$ when $\delta\theta_1 = 0$, $\delta z_0 = 0$, and $\delta\theta_0$ is small. Express your result in terms of the matrices $A_0$, $B_0$, $A_1$, and $B_1$.
—files—

**Q2.3**
8 Points
For the back propagation, give an expression for $\epsilon_k$ so that

$$g_{0,k} = \nabla_\theta \left\{ \|x_k - f_\theta(y_k)\|^2 \right\} .$$

—files—

**Q2.4**
8 Points
For the back propagation, give an expression for the gradient of the total loss function

$$g_0 = \nabla_\theta L(\theta) \ ,$$

in terms of the vectors $g_{0,k}$.
—files—

**Solution:**

**Q2.1**

$$A_0 = \nabla_{z_0} f_{0,\theta_0}(z_0)$$
$$B_0 = \nabla_{\theta_0} f_{0,\theta_0}(z_0)$$
$$A_1 = \nabla_{z_1} f_{1,\theta_1}(z_1)$$
$$B_1 = \nabla_{\theta_1} f_{1,\theta_1}(z_1)$$

**Q2.2**

$$\delta\hat{x} = A_1 \delta z_1 = A_1 B_0 \delta\theta_0$$

**Q2.3**

$$\epsilon_k = -2\left(x_k - f_\theta(y_k)\right)$$

**Q2.4**

Due to typo in exam, the answer was

$$g_0 = \nabla_{\theta_0} L(\theta) = \sum_{k=0}^{K-1} g_{0,k} \ .$$

However, based on classed notes, the answer would be

$$g_0 = \nabla_{\theta_0} L(\theta) = \frac{1}{K} \sum_{k=0}^{K-1} g_{0,k} \ .$$

So either get full credit.

**Q3 Estimation**
34 Points
Consider the forward model with the form

$$X_k = f_\theta(Y_k) + W_k \ ,$$

where $Y_k \sim p(y)$ are i.i.d. vectors for $k = 0, \cdots, K-1$, $f_\theta(\cdot)$ is the inference function parameterized by the vector $\theta$, and $W_k \sim N(0, \sigma^2 I)$ are i.i.d. noise vectors.

**Q3.1**
5 Points
Give an expression for $p_\theta(x|y)$, the conditional density of $X_k$ given $Y_k$ for known $\theta$.
—files—

**Q3.2**
5 Points
Calculate an expression for the maximum likelihood (ML) estimate of $\theta$ given $(X_k, Y_k)$ for $k = 0, \cdots, K-1$.
—files—

**Q3.3**
8 Points
In this subquestion, take a Bayesian approach and assume that $\theta$ is a random vector composed of i.i.d. components, each having an exponential density given by

$$\theta_i \sim g(t) = \frac{1}{2\alpha} \exp\left\{ -\frac{|t|}{\alpha} \right\} \ .$$

Making this new assumption, give an expression for the MAP estimate of $\theta$ given $(X_k, Y_k)$ for $k = 0, \cdots, K-1$.
—files—

**Q3.4**
8 Points
Explain in words, a) the advantages of the MAP estimate over the ML estimate, b) the advantages of the ML estimate over the MAP estimate.
[____]

---

**Solution:**

**Q3.1**

Conditioned on knowledge of $Y_k$, we know that

$$p_\theta(x_k|y_k) = \frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \|x_k - f_\theta(y_k)\|^2 \right\} \ .$$

**Q3.2**

From the previous problem and using the fact that both $Y_k$ and $W_k$ are i.i.d., we know that

$$\begin{aligned} p_\theta(x_0, \cdots, x_{K-1}|y_0, \cdots, y_{K-1}) &= \prod_{k=0}^{K-1} p_\theta(x_k|y_k) \\ &= \prod_{k=0}^{K-1} \frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\|x_k - f_\theta(y_k)\|^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{Kp}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=0}^{K-1}\|x_k - f_\theta(y_k)\|^2\right\} \ . \end{aligned}$$

So then the maximum likelihood (ML) estimate is given by

$$\begin{aligned} \hat{\theta}_{ML} &= \arg\min_\theta \{-\log p_\theta(x|y)\} \\ &= \arg\min_\theta \left\{\frac{1}{2\sigma^2}\sum_{k=0}^{K-1}\|x_k - f_\theta(y_k)\|^2 + \frac{Kp}{2}\log\left(2\pi\sigma^2\right)\right\} \\ &= \arg\min_\theta \left\{\frac{1}{K}\sum_{k=0}^{K-1}\|x_k - f_\theta(y_k)\|^2\right\} \\ &= \arg\min_\theta \{L_{MSE}(\theta)\} \ . \end{aligned}$$

**Q3.3**

In this case, the MAP estimate is given by

$$\begin{aligned} \hat{\theta}_{MAP} &= \arg\min_\theta \{-\log p_\theta(x|y) - \log g(\,\theta\,)\} \\ &= \arg\min_\theta \left\{\frac{1}{2\sigma^2}\sum_{k=0}^{K-1}\|x_k - f_\theta(y_k)\|^2 + \frac{1}{2\alpha}\|\theta\|_1\right\} \\ &= \arg\min_\theta \left\{\frac{1}{K}\sum_{k=0}^{K-1}\|x_k - f_\theta(y_k)\|^2 + \frac{\sigma^2}{\alpha K}\|\theta\|_1\right\} \\ &= \arg\min_\theta \left\{L_{MSE}(\theta) + \frac{\sigma^2}{\alpha K}\|\theta\|_1\right\} \ . \end{aligned}$$

**Q3.4**

a) The advantages of MAP:

- It may result in a lower variance estimates if the prior model is accurate.
- It can generate a more accurate result when the amount of data is small and the number of unknown parameters is large.

b) the advantages of ML:

- It does not require the selection of a prior model.
- It is (mostly) unbiased.
- It is asymptotically efficient.

**Q4 Interpretation of Loss Functions**
32 Points
While training a deep neural network (DNN), you decide to partition your training data into three subsets: the training data - $S_T$; the validation data - $S_V$; and the testing data - $S_E$. These three data subsets are associated with three separate loss functions given by

$$L_T(\theta) = \frac{1}{|S_T|} \sum_{k \in S_T} \|y_k - f_\theta(x_k)\|^2$$

$$L_V(\theta) = \frac{1}{|S_V|} \sum_{k \in S_V} \|y_k - f_\theta(x_k)\|^2$$

$$L_E(\theta) = \frac{1}{|S_E|} \sum_{k \in S_E} \|y_k - f_\theta(x_k)\|^2 \ .$$

where $|S_T|$, $|S_V|$, and $|S_E|$ denote the number training pairs in each subset.

**Q4.1**
8 Points
For this sub problem, assume that after training, $L_V >> L_T$. a) What does this tell you about the capacity of the model? b) What does this tell you about the amount of training data?
—files—

**Q4.2**
8 Points
For this sub problem, assume that after training, $L_V >> L_T$. What options do you have in this case to improve the accuracy of the DNN?
—files—

**Q4.3**
8 Points
For this sub problem, assume that after training, $L_V \approx L_T$. a) What does this tell you about the capacity of the model? b) What does this tell you about the amount of training data?
—files—

**Q4.4**
8 Points
For this sub problem, assume that after training, $L_V \approx L_T$. What options do you have in this case to improve the accuracy of the DNN?
—files—

**Solution:**

**Q4.1**

a) The capacity of the model is high.
b) The amount of training data is insufficient.

**Q4.2**

The approaches to improve the accuracy of DNN:

- Early termination of training.
- Regularization.
- Drop-out method.
- Reduce model order.
- Increase the amount of data used to train the model.

**Q4.3**

a) It is likely that the capacity is low.
b) The amount of data is sufficient for the capacity of the model.

**Q4.4**

The approaches to improve the accuracy of DNN:

- Increase the model order.
- Increase the training time to see if the training loss function, $L_T(\theta)$, can be further minimized.