

---

# Machine Learning

## Answer Sheet for Homework 8

---

Da-Min HUANG

R04942045

*Graduate Institute of Communication Engineering, National Taiwan University*

January 20, 2016

### Problem 1

1. Forward:

$$(A + 1) \times B + (B + 1) \times 1 = (A + 2) B + 1 \quad (1)$$

2. Backward:

$\delta_1^{(L)} = -2 \left( y_n - s_1^{(L)} \right) x_i^{(L-1)}$  counts and

$$\frac{\partial e_n}{\partial w_{ij}^{(\ell)}} = \delta_j^{(\ell)} x_i^{(\ell-1)} \text{ for } 0 \leq i \leq d^{(\ell-1)} \text{ and } 1 \leq j \leq d^{(\ell)} \quad (2)$$

with

$$\delta_j^{(\ell)} = \sum_k \left( \delta_k^{(\ell+1)} \right) \left( w_{jk}^{(\ell+1)} \right) \left( \tanh' \left( s_j^\ell \right) \right) \quad (3)$$

So one backward counts

$$\underbrace{(B + 1) \times 1}_{\text{output layer}} + \underbrace{B \times (A + 1)}_{\text{hidden layer}} + \underbrace{B}_{\text{hidden layer } \delta_j^{(\ell)}} = (A + 3) B + 1 \quad (4)$$

Hence, total number of operations required in a single iteration of backpropagation is

$$((A + 2) B + 1) + ((A + 3) B + 1) = (2A + 5) B + 2 \quad (5)$$

□

## Problem 2

Suppose we have  $k$  hidden layers, which means  $L = k + 1$ , with  $d^{(1)}, d^{(2)}, \dots, d^{(k)}$  units ( $x_0^{(\ell)}$  is not counted here) in each layer. The number of total weights is

$$\sum_{i=0}^{k-1} (d^{(i)} + 1) d^{(i+1)} + (d^{(k)} + 1) \times 1 = \sum_{i=0}^{k-1} d^{(i)} d^{(i+1)} + \sum_{j=1}^k d^{(j)} + (d^{(k)} + 1) := N_w \quad (6)$$

with

$$\sum_{j=1}^k (d^{(j)} + 1) = \left( \sum_{j=1}^k d^{(j)} \right) + k = 36 \text{ and } d^{(0)} = 9 \quad (7)$$

So we have

$$N_w = (37 - k) + 9d^{(1)} + \left( \sum_{i=1}^{k-1} d^{(i)} d^{(i+1)} \right) + d^{(k)} \quad (8)$$

Since  $d^{(\ell)} \geq 1$  for  $0 \leq \ell \leq k + 1$ , so we have  $1 \leq k \leq 18$ .

Claim:  $k = 18$  minimizes  $N_w$ .

Proof of Claim:

If  $k = 18$ , we have 2 units in each hidden layer (one is  $x_0^{(\ell)}$ , not counted in  $d^{(\ell)}$ ), so

$$N_w|_{k=18} = (37 - 18) + 9 \times 1 + \left( \sum_{i=1}^{17} 1 \times 1 \right) + 1 = 46 \quad (9)$$

If  $k = 18 - m$ ,  $m \in \mathbb{N}$  and  $1 \leq m \leq 17$ , we have

$$N_w|_{k=18-m} = (19 + m) + 9d'^{(1)} + \left( \sum_{i=1}^{17-m} d'^{(i)} d'^{(i+1)} \right) + d'^{(18-m)} \quad (10)$$

$$\geq (19 + m) + 9 + (17 - m) + 1 \quad (11)$$

$$= 19 + 9 + 17 + 1 = N_w|_{k=18} \quad (12)$$

where  $d'^{(\ell)}$  is the new number of each hidden layer if  $k = 18 - m$  and (11) holds due to  $d'^{(i)} d'^{(i+1)} \geq 1$  and  $d'^{(i)} \geq 1, \forall i$  (by definition).

Hence, we have  $N_w \geq 46$ .

□

## Problem 3

Following the setting of Problem 2.

Claim:  $k = 2$  with 21 units (not included  $x_0^{(1)}$ ) in  $d^{(1)}$  and 13 units (not included  $x_0^{(2)}$ ) in  $d^{(2)}$  maximizes  $N_w$ .

Proof of Claim:

If  $k = 2$  with 21 units (not included  $x_0^{(1)}$ ) in  $d^{(1)}$  and 13 units (not included  $x_0^{(2)}$ ) in  $d^{(2)}$ , we have

$$N_w|_{k=2} = (37 - 2) + 9 \times 21 + (21 \times 13) + 13 = 510 \quad (13)$$

Consider following cases,

1. If  $k = 2$  with  $34 - m$  units (not included  $x_0^{(1)}$ ) in  $d^{(1)}$  and  $m$  units (not included  $x_0^{(2)}$ ) in  $d^{(2)}$ , where  $m \in \mathbb{N}$  and  $1 \leq m \leq 33$ , we have

$$N_w|_{k=2} = (37 - 2) + 9 \times (34 - m) + ((34 - m) \times m) + m = -(m - 13)^2 + 510 \quad (14)$$

Hence,  $m = 13$  maximize  $N_w|_{k=2}$ .

2. If  $k = 1$ .

$$N_w|_{k=1} = (37 - 1) + 9 \times 35 + 35 = 386 < 510 \quad (15)$$

3. If  $k = 3$  with  $33 - n_1 - n_2$  units (not included  $x_0^{(1)}$ ) in  $d^{(1)}$ ,  $n_1$  units (not included  $x_0^{(2)}$ ) in  $d^{(2)}$  and  $n_2$  units (not included  $x_0^{(3)}$ ) in  $d^{(3)}$ , where  $n_1, n_2 \in \mathbb{N}$  and  $1 \leq n_1, n_2 \leq 32$ , we have

$$N_w|_{k=3} = (37 - 3) + 9 \times (33 - n_1 - n_2) + ((33 - n_1 - n_2) \times n_1 + n_1 \times n_2) + n_2 \quad (16)$$

$$= -(n_1 - 12)^2 - 8n_2 + 475 \leq -8n_2 + 475 \leq 467 < 510 \quad (17)$$

We can see that if we have no  $d^{(3)}$  layer (which means  $n_2 = 0$ ), then  $N_w|_{k=3}$  can be larger.

4. If  $18 \geq k = s \geq 4$  with with  $(36 - s) - \sum_{i=1}^{s-1} n_i$  units (not included  $x_0^{(1)}$ ) in  $d^{(1)}$ ,  $n_1$  units (not included  $x_0^{(2)}$ ) in  $d^{(2)}$ ,  $n_2$  units (not included  $x_0^{(3)}$ ) in  $d^{(3)}$ , ...,  $n_{s-1}$  units (not included  $x_0^{(s)}$ ) in  $d^{(s)}$  where  $n_i \in \mathbb{N}$  and  $1 \leq n_i \leq (35 - s)$ ,  $\forall i$ , we have

$$\begin{aligned} N_w|_{k=s} &= (37 - s) + 9 \times \left( (36 - s) - \sum_{i=1}^{s-1} n_i \right) \\ &\quad + \left( \left( (36 - s) - \sum_{i=1}^{s-1} n_i \right) \times n_1 + \cdots + n_{s-2} \times n_{s-1} \right) + n_{s-1} \end{aligned} \quad (18)$$

We can find that there is no  $n_1 n_2$  term, only  $n_2 n_3$  term exists and no other terms contains  $n_2$ . We have

$$\frac{\partial N_w|_{k=s}}{\partial n_2} = n_3 = 0 \text{ as } N_w|_{k=s} \text{ reaches maximum} \quad (19)$$

This implies  $N_w|_{k=s}$  can be larger without  $d^{(4)}$  layer. So  $k = s \geq 4$  cannot maximize  $N_w$ .

Hence, we have  $N_w \geq 510$ .

□

## Problem 4

$$\nabla_{\mathbf{w}} \text{err}_n(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}} \|\mathbf{x}_n - \mathbf{w} \mathbf{w}^T \mathbf{x}_n\|^2 = \frac{\partial}{\partial \mathbf{w}} (\mathbf{x}_n - \mathbf{w} \mathbf{w}^T \mathbf{x}_n)^T (\mathbf{x}_n - \mathbf{w} \mathbf{w}^T \mathbf{x}_n) \quad (20)$$

$$= \frac{\partial}{\partial \mathbf{w}} (\mathbf{x}_n^T - \mathbf{x}_n^T \mathbf{w}^T \mathbf{w}) (\mathbf{x}_n - \mathbf{w} \mathbf{w}^T \mathbf{x}_n) \quad (21)$$

$$= (-2 \mathbf{x}_n^T \mathbf{w}) (\mathbf{x}_n - \mathbf{w} \mathbf{w}^T \mathbf{x}_n) + (\mathbf{x}_n^T - \mathbf{x}_n^T \mathbf{w}^T \mathbf{w}) (-2 \mathbf{w} \mathbf{x}_n) \quad (22)$$

$$= -2 (\mathbf{x}_n^T \mathbf{w}) \mathbf{x}_n + 2 (\mathbf{x}_n^T \mathbf{w}) \mathbf{w} (\mathbf{x}_n^T \mathbf{w})^T - 2 (\mathbf{x}_n^T \mathbf{w}) \mathbf{x}_n + 2 \mathbf{x}_n^T (\mathbf{w}^T \mathbf{w}) \mathbf{w} \mathbf{x}_n \quad (23)$$

$$= -4 (\mathbf{x}_n^T \mathbf{w}) \mathbf{x}_n + 2 (\mathbf{x}_n^T \mathbf{w})^2 \mathbf{w} + 2 \mathbf{x}_n^T (\mathbf{w}^T \mathbf{w}) \mathbf{w} \mathbf{x}_n \quad (24)$$

where we have used

$$(\mathbf{x}_n^T \mathbf{w}) \mathbf{w} (\mathbf{x}_n^T \mathbf{w})^T = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}^T (\mathbf{x}_n^T \mathbf{w})^T \quad (25)$$

$$= c \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} c = \begin{pmatrix} c^2 w_1 \\ c^2 w_2 \\ \vdots \\ c^2 w_n \end{pmatrix} = (\mathbf{x}_n^T \mathbf{w})^2 \mathbf{w} \quad (26)$$

with  $c := \sum_{i=1}^n x_i w_i$ .

□

### Problem 5

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T (\mathbf{x}_n + \boldsymbol{\epsilon}_n))^T (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T (\mathbf{x}_n + \boldsymbol{\epsilon}_n)) \quad (27)$$

$$= \frac{1}{N} \sum_{n=1}^N \left( \mathbf{x}_n^T - (\mathbf{x}_n + \boldsymbol{\epsilon}_n)^T \mathbf{w}^T \mathbf{w} \right) (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T (\mathbf{x}_n + \boldsymbol{\epsilon}_n)) \quad (28)$$

$$= \frac{1}{N} \sum_{n=1}^N \left\| \mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n \right\|^2 - \boldsymbol{\epsilon}_n^T \mathbf{w}^T \mathbf{w} (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n) \\ - \mathbf{w}\mathbf{w}^T \boldsymbol{\epsilon}_n (\mathbf{x}_n^T - \mathbf{x}_n^T \mathbf{w}^T \mathbf{w}) + (\boldsymbol{\epsilon}_n)^2 (\mathbf{w}^T \mathbf{w})^2 \quad (29)$$

Since  $\boldsymbol{\epsilon}_n$  is generated from a zero-mean, unit variance Gaussian distribution, so  $\mathcal{E}(\boldsymbol{\epsilon}) = 0$  and  $\mathcal{E}(\|\boldsymbol{\epsilon}\|^2) = 1$ . Hence

$$\mathcal{E}(E_{\text{in}}(\mathbf{w})) = \frac{1}{N} \sum_{n=1}^N \left\| \mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n \right\|^2 + (\mathbf{w}^T \mathbf{w})^2 \quad (30)$$

So  $\Omega(\mathbf{w}) = (\mathbf{w}^T \mathbf{w})^2$ .

□

---

### Problem 6

□

---

### Problem 7

□

---

### Problem 8

□

---

### Problem 9

□

**Problem 10**

☐

---

**Problem 11**

☐

---

**Problem 12**

☐

---

**Problem 13**

☐

---

**Problem 14**

☐

---

**Problem 15**

☐

---

**Problem 16**

☐

---

**Problem 17**

☐

### Problem 18



---

### Problem 19



---

### Problem 20



---

## Reference

- [1] Lecture Notes by Hsuan-Tien LIN, Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan.