Derivation of Back Propagation Equations

Steven Brawer

May 2, 2017

notation and definitions

FP = forward propagation
BP = back propagation

s_l - number of nodes in layer l, excluding the bias node.

L - total number of layers. Layers are numbered 1,2,...,L. Layer 1 is input layer, layer L is output layer.

m - number of input samples

n - number of features. For node 1, $s_1 = n$

K = number of classes = s_L

(k,*l*) = node k, layer *l*, is a convenient notation.

$x_1, x_2, ..., x_n$ - Feature (input) values for a particular training-set sample. The sample number i is implicit.

nodes are numbered 0,1,...., s_l . Node 0 is always the bias node, so $x_0 = 1$.

$y_1, y_2, ..., y_K$ - training set output values for a particular sample. Again, sample i is implicit.

z_k^l - **Value of node k, layer *l*.**

- **Each node has a value. It is not necessary to distinguish between an input and output value of a node** (see below eq (8a)).
- **Equations describe all nodes and all layers, including bias nodes.**
- **Sample set index i is always implicit.**

So for input layer 1, for example, for k = 0,1,2,....

$$z_k^1 = x_k \, . \qquad (1)$$

For all bias nodes (node 0 of level *l*)

$$z_0^l = 1 \qquad (1a)$$

The **sigmoid function h(z)** is

$$h(z) \equiv \frac{1}{1+\ell^{-z}} \tag{2}$$

We define the sigmoid function for layer l as

$$h^{l}(z) \equiv \begin{cases} \dfrac{1}{1+\ell^{-z}} & \text{if l} > 1 \\ z & \text{if l} = 1 \end{cases} \tag{3}$$

Note that the function h can be different for every level. The general expression for t**he value of node (j,*l*),** *l* > 1:

$$\boxed{z_{j}^{l} = \begin{cases} x_{j} & \text{l} = 1 \\ \displaystyle\sum_{k=0}^{s_{l-1}} \theta_{j,k}^{(l-1)} h^{(l-1)}\left(z_{k}^{(l-1)}\right) C_{j,k}^{(l-1)} & \text{l=2,3,...,L} \end{cases}} \tag{5}$$

Eq (5) DEFINES the network - it is the definition of FP. This equation is valid for all j, including j = 0 (bias nodes). The θ are weights (see below for a discussion of indices and dimensions). As will be seen, eq(1) is consistent with eqs (5).

Eq (5) is a recursion relation. We define several quantities.

$$\theta_{0,k}^{l} \equiv \begin{cases} 1 & \text{k} = 0 \\ 0 & \text{k} > 0 \end{cases} \tag{6}$$

and

$$C_{j,k}^{l} \equiv \begin{cases} \dfrac{1}{h^{l}(1)} & \text{if k} = 0 \\ 1 & \text{otherwise} \end{cases} \tag{7}$$

Note that for layer 1, $C^1_{j,k}$ = 1 for all j,k. Note that the j index is superfluous, and will be eliminated in a future version of the manuscript.

The Cs will never be used in any actual calculation. Rather, they allow notational simplification. Similarly, $\theta^l_{0,k}$ is never used in an actual calculation. Note that, using (1), (6) and (7), eq (5) gives

$$z^l_0 = 1 \tag{8}$$

for all layers *l*.

Finally, the k=0 term in the sum of eq (5) is the contribution to node j, level *l* from the bias node 0 of layer *l-1.* Using the above definitions and eq (8), this contribution is easily seen to be $\theta^{(l-1)}_{j,0}$. That is,

$$\begin{aligned} z^l_j &= \theta^{(l-1)}_{j,0} h^{(l-1)}\left(z^{(l-1)}_0\right) C^{(l-1)}_{j,0} + \text{ k} > 0 \text{ terms} \\ &= \theta^{(l-1)}_{j,0} + \text{k} > 0 \text{ terms} \end{aligned} \tag{8a}$$

In interpreting these equations, it is useful to keep in mind that the indices of θ are

$$\theta^{l-1}_{\text{index of layer l, index of layer l-1}}$$

That is,

$$\theta^{l-1}_{p,q}$$

"goes" **from** node q, layer *l*-1 **to** node p, layer *l.* Because of (6), in calculations we don't include $\theta^l_{0,q}$ (bias nodes have no input) but we do include $\theta^l_{p,0}$. Therefore, for calculations, $\theta^{l-1}_{p,q}$ has dimension $s_l \times s_{l-1} + 1$.

To interpret eq (5), we could say that the "output" of node (k,l-1) is $h^{(l-1)}\left(z_k^{(l-1)}\right)$. This is sometimes called a_k^{l-1}. (This notation is not needed here) Then eq (5) states that the value of node (j,l) is the sum over "outputs" of node (k, l-1) multiplied by appropriate weights. I prefer the terminology that the value of node (j,l) is the weighted sum over some differentiable function of the value of nodes of (k,l-1). In any event, a variable a_k^{l-1} is superfluous, at least for derivations.

partial derivatives

This section illustrates the use of partial derivatives. Define

$$I_{p,q} = \begin{cases} 1 & \text{if p = q} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

This is the delta-function. However, because the symbol δ is used later in BP, we indicate the delta-function by the symbol I.

The basic derivative is (this is really a total derivative, but we use partial notation)

$$\frac{\partial \theta_{a,b}^{l}}{\partial \theta_{p,q}^{l'}} = I_{a,p} I_{b,q} I_{l,l'} \qquad \mathrm{p} > 0 \tag{10}$$

This equation makes sense only for p > 0. The quantity $\theta_{0,k}^{(l)}$ is a constant (see eq (6)) and so cannot be varied. Therefore, the index a > 0 also is required. However, eq (10) is perfectly reasonable for q = 0.

Now consider from (5)

$$\frac{\partial z_j^l}{\partial \theta_{p,q}^{(l-1)}} = \frac{\partial}{\partial \theta_{p,q}^{(l-1)}} \sum_{k=0}^{s_{l-1}} \theta_{j,k}^{(l-1)} h^{(l-1)}\left(z_k^{(l-1)}\right) C_{j,k}^{(l-1)} \tag{11}$$

The partial derivative works on two terms - θ and h(z). The derivative makes no sense for l=1. Note that z_k^{l-1} depends on $\theta_{k,r}^{(l-2)}$ and so is not affected by this particular derivative. Therefore, for l > 1 (and p > 0)

$$\frac{\partial z_j^l}{\partial \theta_{p,q}^{(l-1)}} = I_{p,j} h^{(l-1)}\left(z_q^{(l-1)}\right) C_{p,q}^{(l-1)} \tag{12}$$

The q = 0 term of (12) is

$$\frac{\partial z_j^l}{\partial \theta_{p,0}^{(l-1)}} = I_{p,j} \tag{13}$$

which is exactly what we expect. The value of a bias node is 1, and its input to the next layer is the weight. Eq (13) is consistent with eqs (8).

For special case *l* = 2, using (1), (3) and (7) in (12)

$$\frac{\partial z_j^2}{\partial \theta_{p,q}^{(1)}} = I_{p,j} x_q \tag{14}$$

which is valid for p>0 and q=0,1,...,n. So the value of the input to a node on level 2 varies linearly with x (outputs from level 1). Note that, because p > 0, we must have j > 0, meaning eq (14) does not apply to the value of the level-2 bias node. That constraint is embodied in the I term. This generalizes to all layers.

Now consider $\frac{\partial z_j^l}{\partial \theta_{p,q}^{(l-2)}}$, where now we differentiate for layer *l* - 2. This only makes sense for *l* >=3. Define the derivative

$$\Psi^{(l)}(z) \equiv \frac{\partial h^{(l)}(z)}{\partial z} \tag{15}$$

We note for future reference that, from (2) and (3),

$$\Psi^{(l)}(z) = \begin{cases} h(z)(1-h(z)) & \text{when } \mathrm{l} > 1 \\ 1 & \text{when } \mathrm{l} = 1 \end{cases} \tag{15a}$$

To illustrate the use of the chain rule for derivatives, and eq (5), we can examine

$$
\begin{aligned}
\frac{\partial z_j^l}{\partial \theta_{p,q}^{(l-2)}} &= \sum_{k=0}^{s_{l-1}} \theta_{j,k}^{(l-1)} C_{j,k}^{(l-1)} \frac{\partial}{\partial \theta_{p,q}^{l-2}} h^{l-1}\left(z_k^{l-1}\right) \\
&= \sum_{k=0}^{s_{l-1}} \theta_{j,k}^{(l-1)} C_{j,k}^{(l-1)} \Psi^{(l-1)}\left(z_k^{l-1}\right) \frac{\partial}{\partial \theta_{p,q}^{l-2}} \sum_{r=0}^{s_l-2} \theta_{k,r}^{(l-2)} h^{(l-2)}\left(z_r^{(l-2)}\right) C_{k,r}^{(l-2)}
\end{aligned} \tag{16}
$$

where in the second line we have used eq (5) for z_k^{l-1}. Because of (10), we must have k = p in the first sum and r = q in the second sum. The quantity $z_r^{(l-2)}$ depends only on $\theta_{r,s}^{(l-3)}$, and so is not affected by the derivative. Therefore

$$
\frac{\partial z_j^l}{\partial \theta_{p,q}^{(l-2)}} = \theta_{j,p}^{(l-1)} C_{j,p}^{(l-1)} \Psi^{(l-1)}\left(z_p^{l-1}\right) h^{(l-2)}\left(z_q^{(l-2)}\right) C_{p,q}^{(l-2)} \tag{17}
$$

Remember, p > 0 in this expression. Parenthetically we can see how our definitions of C, h and so forth simplify the notation of these equations. Without these definitions, we would have to append all sorts of textual constraints on equations such as (17). While this is not substantive, it is notationally convenient.

The interpretation of eq (17) can be illustrated by looking at the q=0 case. The q = 0 case in (17) gives the contribution of the variation in the weight of layer *l*-2 bias node to the variation of the value of node (j, *l*). If q = 0, (17) becomes (for
j > 0)

$$
\frac{\partial z_j^l}{\partial \theta_{p,0}^{(l-2)}} = \theta_{j,p}^{(l-1)} \Psi^{(l-1)}\left(z_p^{l-1}\right) \tag{17a}
$$

This equation can be interpreted as follows. The quantity $\theta_{p,0}^{l-2}$ is the weight *to* node (p,*l*-1) *from* the level *l*-2 bias node. The right side of (17a) is the *derivative* of the output of node (p, *l*-1) times the weight of the contribution of that node to the value of node (j, *l*). If either the weight, or this derivative, were very small, then the contribution of the variation of the weight of the bias node (0, *l* - 2) to node (j, *l*) is a constant, fixed, during the particular BP cycle.

This interpretation is easily generalized to apply to eq (17). This interpretation will be useful in interpreting the so-called error matrix in BP calculations.

cost function and gradients

The cost function J is

$$J \equiv -\frac{1}{m}\sum_{i=1}^{m}\sum_{k_0=1}^{s_L}\left[y\ln\left(h^{(L)}\left(z_{k_0}^{L}\right)\right)+(1-y)\ln\left(1-h^{(L)}\left(z_{k_0}^{L}\right)\right)\right] \tag{18}$$

The sum over i is over all samples. The quantity y is the output for sample i. The dependence of y on i is implicit. The sum over k0 is over output nodes of output layer L - ie, over all K classes. Since there are no bias nodes in layer L, this sum starts at k0=1. The z are also dependent on sample i through their (recursive) dependence on x via eq (5).

Since h < 1 and 1-h < 1, and the ys are non-negative, the logs are negative and the cost function is positive.

We are interested in the gradient of J with respect to some θ. Using (15) and (15a), and noting that none of this makes sense unless L > 1, it is easily seen that

$$\frac{\partial J}{\partial \theta}=\frac{1}{m}\sum_{i=1}^{m}\sum_{k_0=1}^{s_L}\left[\left(h^{L}\left(z_{k_0}^{L}\right)-y\right)\frac{\partial z_{k_0}^{L}}{\partial \theta}\right] \tag{19}$$

Define the layer-L **error**

$$\boxed{\delta_p^L \equiv \left(h^L\left(z_p^L\right) - y\right)\left(1 - I_{p,0}\right)} \qquad (20)$$

That is, δ_0^L is meaningless, so is set to 0. Then (19) is

$$\boxed{\frac{\partial J}{\partial \theta} = \frac{1}{m} \sum_{i=1}^{m} \sum_{k_0=1}^{s_L} \left[\delta_{k_0}^L \frac{\partial z_{k_0}^L}{\partial \theta} \right]} \qquad (21)$$

We will now calculate $\dfrac{\partial z_{k_0}^L}{\partial \theta}$ for particular θ. In general.

$$\frac{\partial z_{k_0}^L}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{k_1=0}^{s_{L-1}} \theta_{k_0,k_1}^{(L-1)} h^{(L-1)}\left(z_{k_1}^{(L-1)}\right) C_{k_0,k_1}^{(L-1)} \qquad (22)$$

Note that the sum starts from k1 = 0, to take the layer L-1 bias node into account.

Consider first

$$\frac{\partial z_{k_0}^L}{\partial \theta_{p,q}^{L-1}} = \frac{\partial}{\partial \theta_{p,q}^{L-1}} \sum_{k_1=0}^{s_{L-1}} \theta_{k_0,k_1}^{(L-1)} h^{(L-1)}\left(z_{k_1}^{(L-1)}\right) C_{k_0,k_1}^{(L-1)} \qquad (23)$$

Taking the derivative and using (10) we get

$$\frac{\partial z_{k_0}^L}{\partial \theta_{p,q}^{L-1}} = I_{p,k_0} h^{(L-1)}\left(z_q^{(L-1)}\right) C_{p,q}^{(L-1)} \qquad (24)$$

Here, p > 0. For q = 0, we have the contribution from layer L-1 bias node,

$$\frac{\partial z_{k_0}^L}{\partial \theta_{p,0}^{L-1}} = I_{p,k_0} \tag{25}$$

the same as (13). The cost function gradient is obtained by putting (24) into (21).

$$\boxed{\frac{\partial J}{\partial \theta_{p,q}^{L-1}} = \frac{1}{m}\sum_{i=1}^{m} \delta_p^L h^{(L-1)}\left(z_q^{(L-1)}\right) C_{p,q}^{(L-1)}} \tag{26}$$

If q = 0 this becomes

$$\frac{\partial J}{\partial \theta_{p,0}^{L-1}} = \frac{1}{m}\sum_{i=1}^{m} \delta_p^L \tag{27}$$

which is the average, over the sample set, of layer-L error for the particular output node p (ie, node (p,L)).

If L=2, eq (26) reduces to simple logistic regression with no hidden layers. Assuming a single class, so K = 1, then k0=1 only. Then (26) becomes simply

$$\frac{\partial J}{\partial \theta_{1,q}^{1}} = \frac{1}{m}\sum_{i=1}^{m} \delta_1^2 x_q \tag{28}$$

Since there are no hidden layers, we can drop the 1 subscript on θ, and the superscript as well, and from (20)

$$\frac{\partial J}{\partial \theta_q} = \frac{1}{m}\sum_{i=1}^{m} \left(h\left(z_1^2\right) - y\right) x_q \tag{29}$$

the usual result.

We now consider the next layer. Applying (5) twice, we have a result equivalent to (16)

$$
\begin{aligned}
\frac{\partial z_{k_0}^{L}}{\partial \theta_{p,q}^{L-2}} &= \frac{\partial}{\partial \theta_{p,q}^{L-2}} \sum_{k_1=0}^{s_{L-1}} \theta_{k_0,k_1}^{(L-1)} h^{(L-1)}\left(z_{k_1}^{(L-1)}\right) C_{k_0,k_1}^{(L-1)} \\
&= \sum_{k_1=0}^{s_{L-1}} \theta_{k_0,k_1}^{(L-1)} C_{k_0,k_1}^{(L-1)} \frac{\partial}{\partial \theta_{p,q}^{L-2}} h^{(L-1)}\left(z_{k_1}^{(L-1)}\right) \\
&= \sum_{k_1=0}^{s_{L-1}} \theta_{k_0,k_1}^{(L-1)} C_{k_0,k_1}^{(L-1)} \Psi^{(L-1)}\left(z_{k_1}^{(L-1)}\right) \frac{\partial}{\partial \theta_{p,q}^{L-2}} \sum_{k_2=0}^{s_{L-2}} \theta_{k_1,k_2}^{(L-2)} h^{(L-2)}\left(z_{k_2}^{(L-2)}\right) C_{k_1,k_2}^{(L-2)}
\end{aligned}
\tag{29a}
$$

Using (10) this is

$$
\frac{\partial z_{k_0}^{L}}{\partial \theta_{p,q}^{L-2}} = \theta_{k_0,p}^{(L-1)} C_{k_0,p}^{(L-1)} \Psi^{(L-1)}\left(z_{p}^{(L-1)}\right) h^{(L-2)}\left(z_{q}^{(L-2)}\right) C_{p,q}^{(L-2)} \tag{30}
$$

Note that

$$
\frac{\partial z_{0}^{L}}{\partial \theta_{p,q}^{L-2}} = 0
$$

because p > 0 and eq (6). This is consistent since there is no bias node (node 0) in level L.

Things are messy, so it is useful to define the quantity

$$
\boxed{M_{j,k}^{\,l} \equiv \theta_{j,k}^{(l)} C_{j,k}^{(l)} \Psi^{(l)}\left(z_{k}^{(l)}\right)} \tag{31}
$$

for *l* = 1,2,...,L-1 (but not *l*=L).We see that

$$
\frac{\partial z_{k_0}^{L}}{\partial \theta_{p,q}^{L-2}} = M_{k_0,p}^{(L-1)} h^{(L-2)}\left(z_{q}^{(L-2)}\right) C_{p,q}^{(L-2)} \tag{32}
$$

We will see that the (0,0) element of M is not used. Putting (32) in (21) we find

$$\frac{\partial J}{\partial \theta_{p,q}^{L-2}} = \frac{1}{m}\sum_{i=1}^{m}\sum_{k_0=1}^{s_L}\left[\delta_{k_0}^{L} M_{k_0,p}^{(L-1)} h^{(L-2)}\left(z_q^{(L-2)}\right) C_{1,q}^{(L-2)}\right] \tag{33}$$

In this expression, we have substituted

$$C_{p,q} = C_{1,q} \tag{33a}$$

without any loss of generality (see eq (7)). We want to do this to keep the p index separated (it is superfluous anyway) when vectorized, so gradients of J are given by a separable matrix, and we can use the simple BP algorithm derived here.

Defining

$$\boxed{\delta_{p}^{L-1} \equiv \sum_{k_0=1}^{s_L} \delta_{k_0}^{L} M_{k_0,p}^{(L-1)}\left(1 - I_{p,0}\right)} \tag{34}$$

we find

$$\boxed{\frac{\partial J}{\partial \theta_{p,q}^{L-2}} = \frac{1}{m}\sum_{i=1}^{m} \delta_{p}^{L-1} h^{(L-2)}\left(z_q^{(L-2)}\right) C_{1,q}^{(L-2)}} \tag{35}$$

From (31) and (6) it is seen that that $M_{k_0,0}^{L-1} = 0$ and therefore the p=0 element of δ is 0.

Everything in (35) except for C depends implicitly on i.

Now for the derivative on the next level. We start with the last line of (29a), applying the chain rule again

$$
\begin{aligned}
\frac{\partial z_{k_0}^{L}}{\partial \theta_{p,q}^{L-3}} &= \sum_{k_1=0}^{s_{L-1}} \theta_{k_0,k_1}^{(L-1)} C_{k_0,k_1}^{(L-1)} \Psi^{(L-1)}\left(z_{k_1}^{(L-1)}\right) \frac{\partial}{\partial \theta_{p,q}^{L-3}} \sum_{k_2=0}^{s_{L-2}} \theta_{k_1,k_2}^{(L-2)} h^{(L-2)}\left(z_{k_2}^{(L-2)}\right) C_{k_1,k_2}^{(L-2)} \\
&= \sum_{k_1=0}^{s_{L-1}} \theta_{k_0,k_1}^{(L-1)} C_{k_0,k_1}^{(L-1)} \Psi^{(L-1)}\left(z_{k_1}^{(L-1)}\right) \sum_{k_2=0}^{s_{L-2}} \theta_{k_1,k_2}^{(L-2)} C_{k_1,k_2}^{(L-2)} \Psi^{(L-2)}\left(z_{k_2}^{(L-2)}\right) \\
&\qquad \times \frac{\partial}{\partial \theta_{p,q}^{L-3}} \sum_{k_3=0}^{s_{L-3}} \theta_{k_2,k_3}^{(L-3)} C_{k_2,k_3}^{(L-3)} h^{(L-3)}\left(z_{k_3}^{(L-3)}\right)
\end{aligned} \tag{36}
$$

Taking the derivative and using (10) and (31) and keeping track of indices, it is straightforward to show

$$
\frac{\partial z_{k_0}^{L}}{\partial \theta_{p,q}^{L-3}} = \sum_{k_1=1}^{s_{L-1}} M_{k_0,k_1}^{(L-1)} M_{k_1,p}^{(L-2)} C_{1,q}^{(L-3)} h^{(L-3)}\left(z_q^{(L-3)}\right) \tag{37}
$$

Because k0 >= 1 in the cost function, it is seen from (31) and (6) that the k1=0 term in the sum is 0, so the sum starts from k1=1. With this, the L-3 gradient of the cost function is

$$
\begin{aligned}
\frac{\partial J}{\partial \theta_{p,q}^{L-3}} &= \frac{1}{m} \sum_{i=1}^{m} \sum_{k_0=0}^{s_L} \sum_{k_1=1}^{s_{L-1}} \delta_{k_0}^{L} M_{k_0,k_1}^{(L-1)} M_{k_1,p}^{(L-2)} C_{1,q}^{(L-3)} h^{(L-3)}\left(z_q^{(L-3)}\right) \\
&= \frac{1}{m} \sum_{i=1}^{m} \sum_{k_1=1}^{s_{L-1}} \delta_{k_1}^{L-1} M_{k_1,p}^{(L-2)} C_{1,q}^{(L-3)} h^{(L-3)}\left(z_q^{(L-3)}\right)
\end{aligned} \tag{39}
$$

Define

$$
\boxed{\delta_p^{L-2} \equiv \sum_{k_1=1}^{s_{L-1}} \delta_{k_1}^{L-1} M_{k_1,p}^{(L-2)} \left(1 - I_{p,0}\right)} \tag{40}
$$

Then the derivative becomes

$$\boxed{\frac{\partial J}{\partial \theta_{p,q}^{L-3}}=\frac{1}{m}\sum_{i=0}^{m} \delta_p^{L-2} C_{1,q}^{(L-3)} h^{(L-3)}\left(z_q^{(L-3)}\right)} \tag{41}$$

the general equations

We can keep going to get the generalization of the above equations. The general equations are:

$$\delta_p^L \equiv \left(h^L\left(z_p^L\right)-y\right)\left(1-I_{p,0}\right) \tag{42}$$

For *l* < L, we substitute for M. Since, as discussed, the subscripts on M are always >= 1, The C term is 1 and we can drop it. This gives

$$\delta_p^l \equiv \sum_{k=1}^{s_{l+1}} \delta_k^{l+1} \theta_{k,p}^l \Psi^l\left(z_p^l\right)\left(1-I_{p,0}\right) \tag{43}$$

$$\frac{\partial J}{\partial \theta_{p,q}^{l}}=\frac{1}{m}\sum_{i=0}^{m} \delta_p^{l+1} C_{1,q}^{(l)} h^{(l)}\left(z_q^{(l)}\right) \tag{44}$$

Note that for level 1, from eqs (15), $\Psi^1(z)=1$, and using (3), (7) in (44)

$$h^1\left(z_q^1\right)C_{1,q}^1 = x_q \tag{45}$$

valid for all q (= 0,1,2,...,n). We note that

$$\frac{\partial J}{\partial \theta_{p,0}^{l}} = \frac{1}{m}\sum_{i=0}^{m} \delta_{p}^{l+1} \qquad (46)$$

the average over the sample set of layer *l*+1 error for node p. Compare eq (27).

discussion, dimensions and vectorization

As noted previously, the matrix $\theta_{p,q}^{l-1}$ has dimension $s_l \times s_{l-1} + 1$ (since a p=0 element does not exist, but the q = 0 element does exist). The column $\theta_{p,0}^{l-1}$ has the weights from the bias node. Note that in eq (43), the p=0 column is not used, and δ_p^l has dimension s_l . In vectorizing (43), one must be careful with the 0th column of θ.

To vectorize (43), note that if δ is considered a column vector, then

$$\sum_{k=1} \delta_k^{l+1}\theta_{k,p}^{l} = \left(\theta^{(l)T}\delta^{l+1}\right)_p \qquad (47)$$

the pth element of a column vector, where p > 0. Then considering Ψ as a p-dimensional column vector, we have

$$\delta^l = \left(\theta^{(l)T}\delta^{l+1}\right).*\Psi^l\left(z^l\right) \qquad (48)$$

Note that the weight of the bias node enters the gradient equations only through h and Ψ . The reason is that the input to a node level *l*, say, is linear in the weight from the bias node from level *l* - 1.