

# 后向传递算法

emacsun

## 目录

<b>1 Cost Function</b>	<b>1</b>
<b>2 Backpropagation Algorithm</b>	<b>2</b>
2.1 Back propagation Algorithm . . . . .	3
<b>3 Backpropagation Intuition</b>	<b>4</b>
<b>4 Implementation Note: Unrolling Parameters</b>	<b>5</b>
<b>5 Gradient Checking</b>	<b>6</b>
<b>6 Random Initialization</b>	<b>7</b>
<b>7 Putting it Together</b>	<b>8</b>
7.1 training a Neural Network . . . . .	8

## 1 Cost Function

Let's first define a few variables that we will need to use:

1.  $L$  = total number of layers in the network
2.  $s_l$  = number of units (not counting bias unit) in layer  $l$
3.  $K$  = number of output units/classes

Recall that in neural networks, we may have many output nodes. We denote  $h_{\Theta}(x)_k$  as being a hypothesis that results in the  $k_{th}$  output. Our cost function



## 2 BACKPROPAGATION ALGORITHM

for neural networks is going to be a generalization of the one we used for logistic regression. Recall that the cost function for regularized logistic regression was:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log(h_{\theta}(x^{(i)})) + (1-y^{(i)}) \log(1-h_{\theta}(x^{(i)}))] + \frac{\lambda}{2m} \sum_{j=1}^n \theta_j^2 \quad (1.1)$$

For neural networks, it is going to be slightly more complicated:

$$J(\Theta) = -\frac{1}{m} \sum_{i=1}^m \sum_{k=1}^K \left[ y_k^{(i)} \log((h_{\Theta}(x^{(i)}))_k) + (1 - y_k^{(i)}) \log(1 - (h_{\Theta}(x^{(i)}))_k) \right] + \frac{\lambda}{2m} \sum_{l=1}^{L-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (\Theta_{j,i}^{(l)})^2$$

We have added a few nested summations to account for our multiple output nodes. In the first part of the equation, before the square brackets, we have an additional nested summation that loops through the number of output nodes.

In the regularization part, after the square brackets, we must account for multiple theta matrices. The number of columns in our current theta matrix is equal to the number of nodes in our current layer (including the bias unit). The number of rows in our current theta matrix is equal to the number of nodes in the next layer (excluding the bias unit). As before with logistic regression, we square every term.

Note:

1. the double sum simply adds up the logistic regression costs calculated for each cell in the output layer
2. the triple sum simply adds up the squares of all the individual  $\Theta$ s in the entire network.
3. the  $i$  in the triple sum does not refer to training example  $i$

## 2 Backpropagation Algorithm

”Backpropagation” is neural-network terminology for minimizing our cost function, just like what we were doing with gradient descent in logistic and linear regression. Our goal is to compute:

$$\min_{\Theta} J(\Theta) \quad (2.1)$$



## 2 BACKPROPAGATION ALGORITHM

That is, we want to minimize our cost function  $J$  using an optimal set of parameters in  $\Theta$ . In this section we'll look at the equations we use to compute the partial derivative of  $J(\Theta)$ :

$$\frac{\partial}{\partial \Theta_{i,j}^{(l)}} J(\Theta) \quad (2.2)$$

To do so, we use the following algorithm:

**Backpropagation algorithm**

- Training set  $\{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$
- Set  $\Delta_{i,j}^{(l)} = 0$  (for all  $l, i, j$ ). *(use to compute  $\frac{\partial}{\partial \Theta_{i,j}^{(l)}} J(\Theta)$ )*
- For  $i = 1$  to  $m \leftarrow (x^{(i)}, y^{(i)})$ .
  - Set  $a^{(1)} = x^{(i)}$
  - Perform forward propagation to compute  $a^{(l)}$  for  $l = 2, 3, \dots, L$
  - Using  $y^{(i)}$ , compute  $\delta^{(L)} = a^{(L)} - y^{(i)}$
  - Compute  $\delta^{(L-1)}, \delta^{(L-2)}, \dots, \delta^{(2)}$   ~~$\delta^{(1)}$~~
  - $\Delta_{i,j}^{(l)} := \Delta_{i,j}^{(l)} + a_j^{(l)} \delta_i^{(l+1)}$   *$\Delta^{(l)} := \Delta^{(l)} + \delta^{(l+1)} (a^{(l)})^T$*
- $D_{i,j}^{(l)} := \frac{1}{m} \Delta_{i,j}^{(l)} + \lambda \Theta_{i,j}^{(l)}$  if  $j \neq 0$
- $D_{i,j}^{(l)} := \frac{1}{m} \Delta_{i,j}^{(l)}$  if  $j = 0$

$\frac{\partial}{\partial \Theta_{i,j}^{(l)}} J(\Theta) = D_{i,j}^{(l)}$

### 2.1 Back propagation Algorithm

Given training set  $\{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$

Set  $\Delta_{i,j}^{(l)} := 0$  for all  $(l, i, j)$ , hence you end up with a matrix full of zeros:

For training example  $t = 1$  to  $m$ :

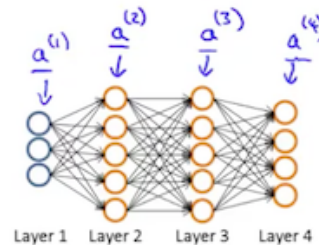
1. set  $a^{(1)} := x^{(t)}$
2. Perform forward propagation to compute to compute  $a^{(l)}$ , for  $l = 2, 3, \dots, L$

#### Gradient computation

Given one training example  $(x, y)$ :

Forward propagation:

$$\begin{aligned}
 & a^{(1)} = x \\
 \rightarrow & z^{(2)} = \Theta^{(1)} a^{(1)} \\
 \rightarrow & a^{(2)} = g(z^{(2)}) \quad (\text{add } a_0^{(2)}) \\
 \rightarrow & z^{(3)} = \Theta^{(2)} a^{(2)} \\
 \rightarrow & a^{(3)} = g(z^{(3)}) \quad (\text{add } a_0^{(3)}) \\
 \rightarrow & z^{(4)} = \Theta^{(3)} a^{(3)} \\
 \rightarrow & a^{(4)} = h_{\Theta}(x) = g(z^{(4)})
 \end{aligned}$$





Using  $y^{(t)}$ , compute  $\delta^{(L)} = a^{(L)} - y^{(t)}$

Where  $L$  is our total number of layers and  $a^{(L)}$  is the vector of outputs of the activation units for the last layer. So our "error values" for the last layer are simply the differences of our actual results in the last layer and the correct outputs in  $y$ . To get the delta values of the layers before the last layer, we can use an equation that steps up back from right to left:

Compute  $\delta^{(L-1)}, \delta^{(L-1)}, \dots, \delta^{(2)}$  using  $\delta^{(l)} = ((\Theta^{(l)})^T \delta^{(l+1)} * a^{(l)} * (1 - a^{(l)}))$

The delta values of layer  $l$  are calculated by multiplying the delta values in the next layer with the theta matrix of layer  $l$ . We then element-wise multiply that with a function called  $g'$ , which is the derivative of the activation function  $g$  evaluated with the input values given by  $z^{(l)}$ .

The  $g'$  derivative terms can also be written out as:

$$g'(z^{(l)}) = a^{(l)} * (1 - a^{(l)}) \quad (2.3)$$

$\Delta_{i,j}^{(l)} := \Delta_{i,j}^{(l)} + a_j^{(l)} \delta_i^{(l+1)}$  or with vectorization,  $\Delta^{(l)} := \Delta^{(l)} + \delta^{(l+1)} (a^{(l)})^T$

Hence we can update our new  $\Delta$  matrix.:

$$\begin{aligned} D_{i,j}^{(l)} &= \frac{1}{m} (\Delta_{i,j}^{(l)} + \lambda \Theta_{i,j}^{(l)}) \quad j \neq 0 \\ D_{i,j}^{(l)} &= \frac{1}{m} \Delta_{i,j}^{(l)} \quad j = 0 \end{aligned}$$

The capital-delta matrix  $D$  is used as an "accumulator" to add up our values as we go along and eventually compute our partial derivative. Thus we get  $\frac{\partial}{\partial \Theta_{i,j}^{(l)}} J(\Theta) = D_{i,j}^{(l)}$

### 3 Backpropagation Intuition

Recall that the cost function for a neural network is :

$$J(\Theta) = -\frac{1}{m} \sum_{t=1}^m \sum_{k=1}^K \left[ y_k^{(t)} \log(h_{\Theta}(x^{(t)}))_k + (1 - y_k^{(t)}) \log(1 - h_{\Theta}(x^{(t)}))_k \right] + \frac{\lambda}{2m} \sum_{l=1}^{L-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (\Theta_{j,i}^{(l)})^2$$

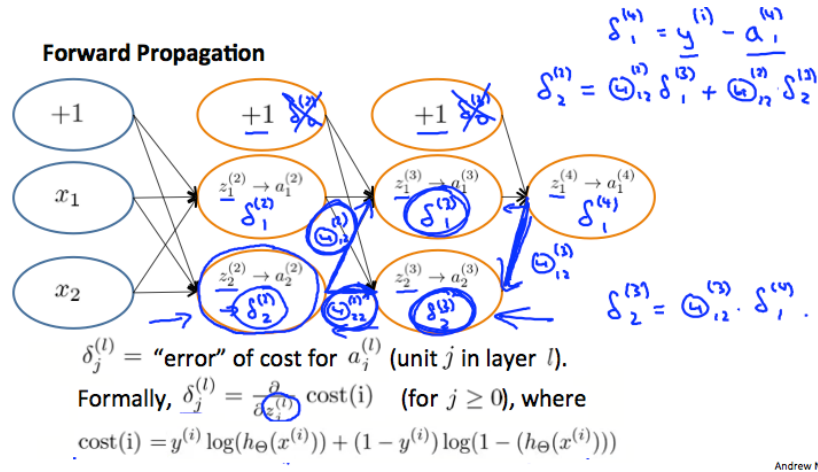
If we consider simple non-multiclass classification ( $k=1$ ) and disregard regularization, the cost is computed with:

$$\text{cost}(t) = y^{(t)} \log(h_{\Theta}(x^{(t)})) + (1 - y^{(t)}) \log(1 - h_{\Theta}(x^{(t)})) \quad (3.1)$$



#### 4 IMPLEMENTATION NOTE: UNROLLING PARAMETERS

Recall that our derivative is the slope of a line tangent to the cost function, so the steeper the slope the more incorrect we are. Let us consider the following neural network below and see how we could calculate some  $\delta_j^{(l)}$ :



In the image above, to calculate  $\delta_2^{(2)}$ , we multiply the weights  $\Theta_{12}^{(2)}$  and  $\Theta_{22}^{(2)}$  by their respective  $\delta$  values found to the right of each edge. So we get  $\delta_2^{(2)} = \Theta_{12}^{(2)} * \delta_1^{(3)} + \Theta_{22}^{(2)} * \delta_2^{(3)}$ . To calculate every single possible  $\delta_j^{(l)}$ , we could start from the right of our diagram. We can think of our edges as our  $\Theta_{ij}$ . Going from right to left, to calculate the value of  $\delta_j^{(l)}$ , you can just take the over all sum of each weight times the  $\delta$  it is coming from. Hence, another example would be  $\delta_2^{(3)} = \Theta_{12}^{(3)} * \delta_1^{(4)}$

## 4 Implementation Note: Unrolling Parameters

With neural networks, we are working with sets of matrices:

$$\Theta^{(1)}, \Theta^{(2)}, \Theta^{(3)}, \dots$$

$$D^{(1)}, D^{(2)}, D^{(3)}, \dots$$

In order to use optimizing functions such as "fminunc()", we will want to "unroll" all the elements and put them into one long vector:

```
thetaVector = [ Theta1(:); Theta2(:); Theta3(:); ]
deltaVector = [ D1(:); D2(:); D3(:) ]
```



If the dimensions of Theta1 is  $10 \times 11$ , Theta2 is  $10 \times 11$  and Theta3 is  $1 \times 11$ , then we can get back our original matrices from the "unrolled" versions as follows:

```
Theta1 = reshape(thetaVector(1:110),10,11)
Theta2 = reshape(thetaVector(111:220),10,11)
Theta3 = reshape(thetaVector(221:231),1,11)
```

To summarize:

#### Learning Algorithm

- Have initial parameters  $\Theta^{(1)}, \Theta^{(2)}, \Theta^{(3)}$ .
- Unroll to get `initialTheta` to pass to
- `fminunc(@costFunction, initialTheta, options)`

```
function [jval, gradientVec] = costFunction(thetaVec)
    From thetaVec, get  $\Theta^{(1)}, \Theta^{(2)}, \Theta^{(3)}$ .
    Use forward prop/back prop to compute  $D^{(1)}, D^{(2)}, D^{(3)}$  and  $J(\Theta)$ .
    Unroll  $D^{(1)}, D^{(2)}, D^{(3)}$  to get gradientVec.
```

## 5 Gradient Checking

Gradient checking will assure that our backpropagation works as intended. We can approximate the derivative of our cost function with:

$$\frac{\partial}{\partial \Theta} J(\Theta) \approx \frac{J(\Theta + \epsilon) - J(\Theta - \epsilon)}{2\epsilon} \quad (5.1)$$

With multiple theta matrices, we can approximate the derivative with respect to  $\Theta_j$  as follows:

$$\frac{\partial}{\partial \Theta_j} J(\Theta) \approx \frac{J(\Theta_1, \dots, \Theta_j + \epsilon, \dots, \Theta_n) - J(\Theta_1, \dots, \Theta_j - \epsilon, \dots, \Theta_n)}{2\epsilon} \quad (5.2)$$

A small value for  $\epsilon$  such as  $\epsilon = 10^{-4}$ , guarantees that the math works out properly. If the value for  $\epsilon$  is too small, we can end up with numerical problems. Hence, we are only adding or subtracting epsilon to the  $\Theta_j$  matrix. In octave we can do it as follows:

```
epsilon = 1e-4;
for i = 1:n,
    thetaPlus = theta;
    thetaPlus(i) += epsilon;
    thetaMinus = theta;
```



```
thetaMinus(i) -= epsilon;  
gradApprox(i) = (J(thetaPlus) - J(thetaMinus))/(2*epsilon)  
end;
```

We previously saw how to calculate the deltaVector. So once we compute our gradApprox vector, we can check that  $\text{gradApprox} \approx \text{deltaVector}$ .

Once you have verified once that your backpropagation algorithm is correct, you don't need to compute gradApprox again. The code to compute gradApprox can be very slow.

## 6 Random Initialization

Initializing all theta weights to zero does not work with neural networks. When we backpropagate, all nodes will update to the same value repeatedly. Instead we can randomly initialize our weights for our  $\Theta$  matrices using the following method:

**Random initialization: Symmetry breaking**

→ Initialize each  $\Theta_{ij}^{(l)}$  to a random value in  $[-\epsilon, \epsilon]$   
(i.e.  $-\epsilon \leq \Theta_{ij}^{(l)} \leq \epsilon$ )

E.g.

→ **Theta1** = `rand(10,11)*(2*INIT_EPSILON) - INIT_EPSILON;` Random 10x11 matrix (betw. 0 and 1)  $[-\epsilon, \epsilon]$

→ **Theta2** = `rand(1,11)*(2*INIT_EPSILON) - INIT_EPSILON;`

Hence, we initialize each  $\Theta_{i,j}^{(l)}$  to a random value between  $[-\epsilon, +\epsilon]$ . Using the above formula guarantees that we get the desired bound. The same procedure applies to all the  $\Theta$ 's. Below is some working code you could use to experiment.

```
If the dimensions of Theta1 is 10x11, Theta2 is 10x11 and Theta3 is 1x11  
Theta1 = rand(10,11) * (2 * INIT_EPSILON) - INIT_EPSILON;  
Theta2 = rand(10,11) * (2 * INIT_EPSILON) - INIT_EPSILON;  
Theta3 = rand(1,11) * (2 * INIT_EPSILON) - INIT_EPSILON;
```



`rand(x,y)` is just a function in octave that will initialize a matrix of random real numbers between 0 and 1.

(Note: the epsilon used above is unrelated to the epsilon from Gradient Checking)

## 7 Putting it Together

First, pick a network architecture; choose the layout of your neural network, including how many hidden units in each layer and how many layers in total you want to have.

1. Number of input units = dimension of features;
2. Number of output units = number of classes;
3. Number of hidden units per layer = usually more the better (must balance with cost of computation as it increase with more hidden units)
4. Defaults: 1 hidden layer. If you have more than 1 hidden layer, then it is recommended that you have the same number of units in every hidden layer.

### 7.1 training a Neural Network

1. Randomly initialize the weights
2. Implement forward propagation to get  $H_{\Theta}(x^{(i)})$  for any  $x^{(i)}$
3. Implement the cost function.
4. Implement the backpropagation to compute partial derivatives
5. Use gradient checking to confirm that your backpropagation works. Then disable gradient checking.
6. Use gradient descent or a built-in optimization function to minimize the cost function with the weights in theta.

When we perform forward and back propagation, we loop on every training example:

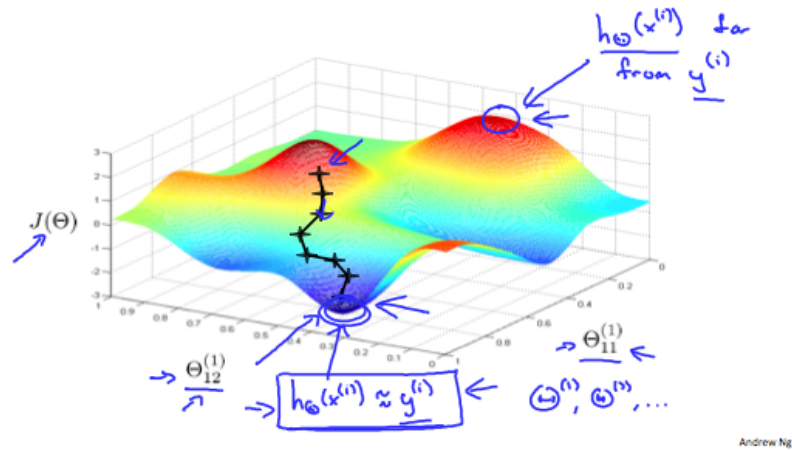




## 7 PUTTING IT TOGETHER

```
for i = 1:m,
    Perform forward propagation and backpropagation using example (x(i),y(i))
    (Get activations a(l) and delta terms d(l) for l = 2,...,L)
```

The following image gives us an intuition of what is happening as we are implementing our neural network:



Ideally, you want  $H_{\Theta}(x^{(i)}) \approx y^{(i)}$ . This will minimize our cost function. However, keep in mind that  $J(\Theta)$  is not convex and thus we can end up in a local minimum instead.