

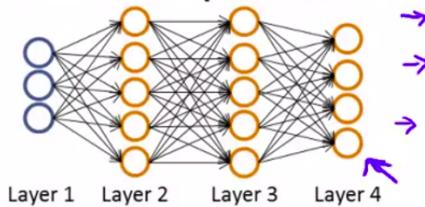
Week 5 - Lecture 1

Neural Networks

Cost Function

Notation:

Neural Network (Classification)



- $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$
- $L = \text{total no. of layers in network}$ $L=4$
- $s_l = \text{no. of units (not counting bias unit) in layer } l$ $s_1=3, s_2=5, s_3=s_L=4$

Binary classification

$$y = 0 \text{ or } 1$$

1 output unit

$$h_{\Theta}(x) \in \mathbb{R}$$

$$s_L = 1, \quad K=1$$

$$h_{\Theta}(x)$$

Multi-class classification (K classes)

$$y \in \mathbb{R}^K \quad \text{E.g. } \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \leftarrow \begin{array}{l} \text{pedestrian} \\ \text{car} \\ \text{motorcycle} \\ \text{truck} \end{array}$$

K output units

$$h_{\Theta}(x) \in \mathbb{R}^k$$

$$s_L = K \quad (K \geq 3)$$

Andrew Ng

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 for neural networks is going to be a generalization of the one we used for logistic regression. Remember:

Logistic Regression Cost Function:

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

For Neural Networks, our cost function will be:

$$h_{\Theta}(x) \in \mathbb{R}^K, \text{ where } (h_{\Theta}(x))_i = i^{th} \text{ output}$$

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

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:

- the double sum simply adds up the logistic regression costs calculated for each cell in the output layer.
- the triple sum simply adds up the squares of all the individual θ^j 's in the entire network.
- the i in the triple sum does not refer to training example i .



Backpropagation Algorithm

Given the cost function $J(\Theta)$, our objective is to compute $\Theta = \arg \min J(\Theta)$. To calculate this value of Θ , we need to compute two things:

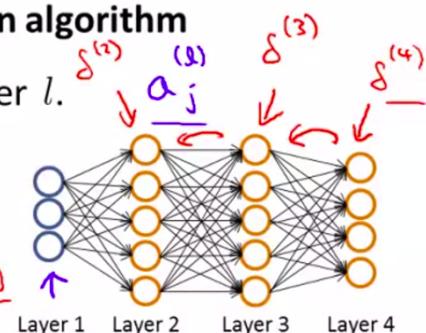
1. The value of $J(\Theta)$.
2. Each partial derivative $\frac{\partial}{\partial \Theta_{ij}^{(l)}} J(\Theta)$.

"Backpropagation" is neural-network terminology for minimizing our cost function.

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 $J(\Theta)$.

Gradient computation: Backpropagation algorithm

Intuition: $\delta_j^{(l)}$ = "error" of node j in layer l .



For each output unit (layer $L = 4$)

$$\delta_j^{(4)} = a_j^{(4)} - y_j$$

$$(h_\Theta(x))_j$$

$$\delta^{(4)} = a^{(4)} - y$$

Layer 1 Layer 2 Layer 3 Layer 4

$$\rightarrow \delta^{(3)} = (\Theta^{(3)})^T \delta^{(4)} \cdot g'(z^{(3)})$$

$$\rightarrow \delta^{(2)} = (\Theta^{(2)})^T \delta^{(3)} \cdot g'(z^{(2)})$$

$$(\text{No } \delta^{(1)})$$

$$a^{(3)} \cdot (1 - a^{(3)})$$

$$a^{(2)} \cdot (1 - a^{(2)})$$

$$\frac{\partial}{\partial \Theta_{ij}^{(k)}} J(\Theta) = a_j^{(k)} \delta_i^{(k+1)}$$

(ignoring λ ; if $\lambda = 0$). ↵

Backpropagation algorithm

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

Algorithm:

Given training set $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}, \dots, (x^{(m)}, y^{(m)})\}$.

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

For training example $t = 1$ to m :

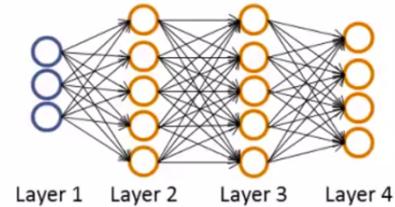
1. Set $a^{(1)} := x^{(i)}$
2. Perform forward propagation to compute $a^{(l)}$ for $l = 2, 3, \dots, L$.

Gradient computation

Given one training example $(\underline{x}, \underline{y})$:

Forward propagation:

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



3. Using y^i , compute $\delta^{(L)} = a^{(L)} - \underline{y}^{(i)}$

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 us back from right to left:

$$4. \text{ Compute } \delta^{L-1}, \delta^{L-2}, \dots, \delta^2 \text{ using } \delta^{(l)} = \left[(\Theta^{(l)})^T \delta^{(l+1)} \right] . * a^{(l)} . * (1 - a^{(l)})$$

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' , or g -prime, which is the derivative of the activation function g evaluated with the input values given by $z^{(l)}$.

The g -prime derivative terms can also be written out as: $g' (z^{(l)}) = a^{(l)} . * (1 - a^{(l)})$

$$5. \Delta_{i,j}^{(l)} := \Delta_{i,j}^{(l)} + a_j^{(l)} \delta_i^{(l+1)}, \text{ or with vectorization } \Delta^{(l)} := \Delta^{(l)} + \delta^{(l+1)} (a^{(l)})^T$$

Hence we update our new Δ matrix.

- $D_{i,j}^{(l)} := \frac{1}{m} (\Delta_{i,j}^{(l)} + \lambda \Theta_{i,j}^{(l)}), \text{ if } j \neq 0.$
- $D_{i,j}^{(l)} := \frac{1}{m} \Delta_{i,j}^{(l)}, \text{ if } j = 0.$

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 $\frac{\partial}{\partial \Theta_{ij}^{(l)}} J(\Theta) = D_{ij}^{(l)}$.

Backpropagation Intuition

Recall that the cost function for a neural network is:

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

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

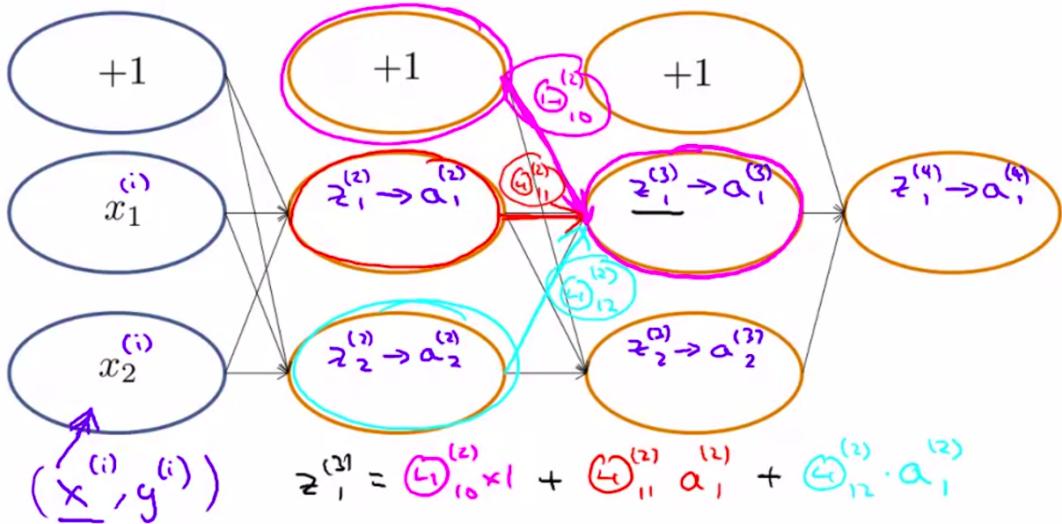
$$\text{cost}(t) = y^{(t)} \log(h_\Theta(x^{(t)})) + (1 - y^{(t)}) \log(1 - h_\Theta(x^{(t)}))$$

Intuitively, $\delta_j^{(l)}$ is the "error" for $a_j^{(l)}$ (unit j in layer l). More formally, the delta values are actually the derivative of the cost function:

$$\delta_j^{(l)} = \frac{\partial}{\partial z_j^{(l)}} \text{cost}(t)$$

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 $\delta_j^{(l)}$:

Forward Propagation



Andrew Ng

In the image above, to calculate $\delta_2^{(2)}$, we multiply the weights $\Theta_{12}^{(2)}$ and $\Theta_{22}^{(2)}$ by their respective δ 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 Θ_{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 δ it is coming from. Hence, another example would be $\delta_2^{(3)} = \Theta_{12}^{(3)} * \delta_1^{(4)}$.

Backpropagation in Practice

Implementation note: Unrolling parameters

Note: this is just relevant, in this course, for Matlab/Octave implementation. Feel free to skip this subsection if you feel it's unnecessary.

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 10×11 , Theta2 is 10×11 and Theta3 is 1×11 , 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)`

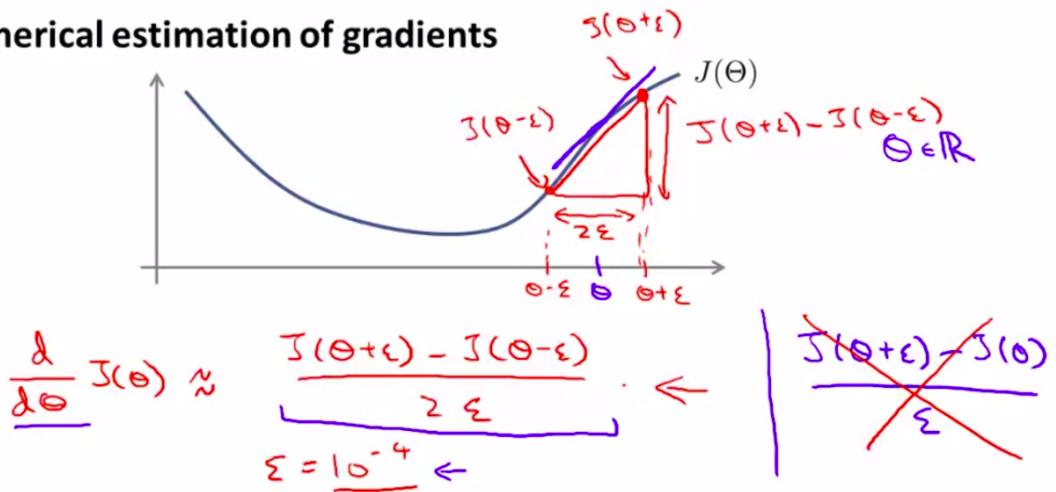
```
function [jval, gradientVec] = costFunction(thetaVec)
    From thetaVec, get  $\Theta^{(1)}, \Theta^{(2)}, \Theta^{(3)}$ .
    Use forward prop/back prop to compute  $D^{(1)}, D^{(2)}, D^{(3)}$  and  $J(\Theta)$ .
    Unroll  $D^{(1)}, D^{(2)}, D^{(3)}$  to get gradientVec.
```

Gradient Checking

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

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

Numerical estimation of gradients



Implement: `gradApprox = (J(theta + EPSILON) - J(theta - EPSILON)) / (2*EPSILON)`

Andrew Ng

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

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

A small value for ϵ (epsilon) such as $\epsilon = 10^{-4}$, guarantees that the math works out properly. If the value for ϵ is too small, we can end up with numerical problems.

Hence, we are only adding or subtracting epsilon to the Θ_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.**

Implementation Note:

- Implement backprop to compute **DVec** (unrolled $D^{(1)}, D^{(2)}, D^{(3)}$).
- Implement numerical gradient check to compute **gradApprox**.
- Make sure they give similar values.
- Turn off gradient checking. Using backprop code for learning.

Important:

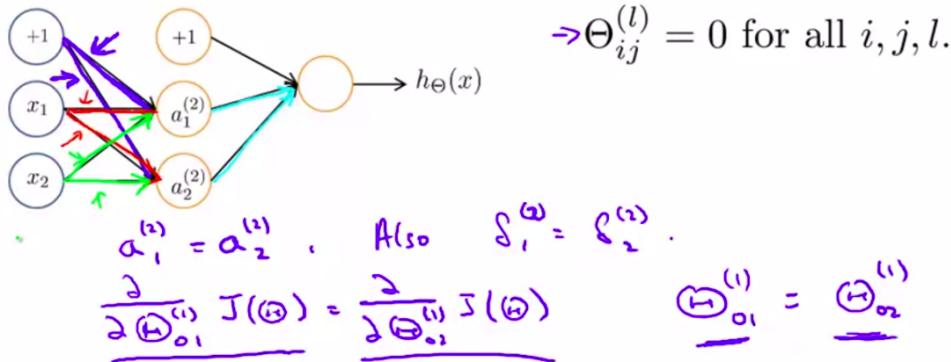
- Be sure to disable your gradient checking code before training your classifier. If you run numerical gradient computation on every iteration of gradient descent (or in the inner loop of **costFunction(...)**) your code will be very slow.

Andrew Ng

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.

Zero initialization



After each update, parameters corresponding to inputs going into each of two hidden units are identical.

Andrew Ng

Instead we can randomly initialize our weights for our Θ 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. Random 10×11 matrix (betw. 0 and 1)
- $\Theta_{\text{eta1}} = \text{rand}(10, 11) * (2 * \text{INIT_EPSILON}) - \text{INIT_EPSILON};$ $[-\epsilon, \epsilon]$
- $\Theta_{\text{eta2}} = \text{rand}(1, 11) * (2 * \text{INIT_EPSILON}) - \text{INIT_EPSILON};$

Hence, we initialize each $\Theta_{ij}^{(l)}$ 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 Θ 's. Below is some working code you could use to experiment.

If the dimensions of Θ_{eta1} is 10×11 , Θ_{eta2} is 10×11 and Θ_{eta3} is 1×11 .

```
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;
```

$\text{rand}(x, y)$ 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)

Putting it Together

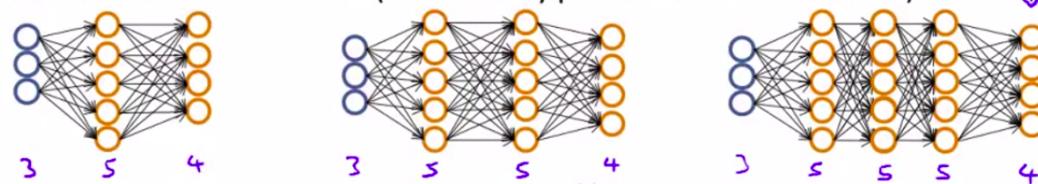
First, pick a network architecture; choose the layout of your neural network, including how many hidden units in each layer and how many layers in total you want to have.

- Number of input units = dimension of features $x^{(i)}$
- Number of output units = number of classes
- Number of hidden units per layer = usually more the better (must balance with cost of computation as it increases with more hidden units)

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.

Training a neural network

Pick a network architecture (connectivity pattern between neurons)



→ No. of input units: Dimension of features $x^{(i)}$

→ No. output units: Number of classes

Reasonable default: 1 hidden layer, or if >1 hidden layer, have same no. of hidden units in every layer (usually the more the better)

$$y \in \{1, 2, 3, \dots, 10\}$$

~~$y \neq 5$~~

$$y = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ or } \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow$$

Andrew Ng

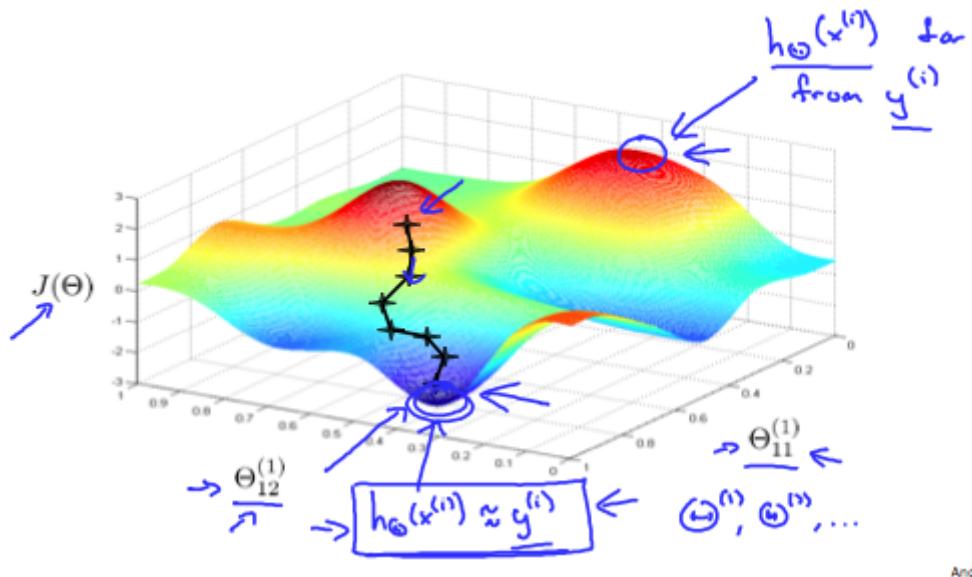
Training a Neural Network

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 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:

```
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:



Andrew Ng

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.

References

- [1] Machine Learning - Stanford University (<https://www.coursera.org/learn/machine-learning>).