BACK PROPAGATION EXPLAINED

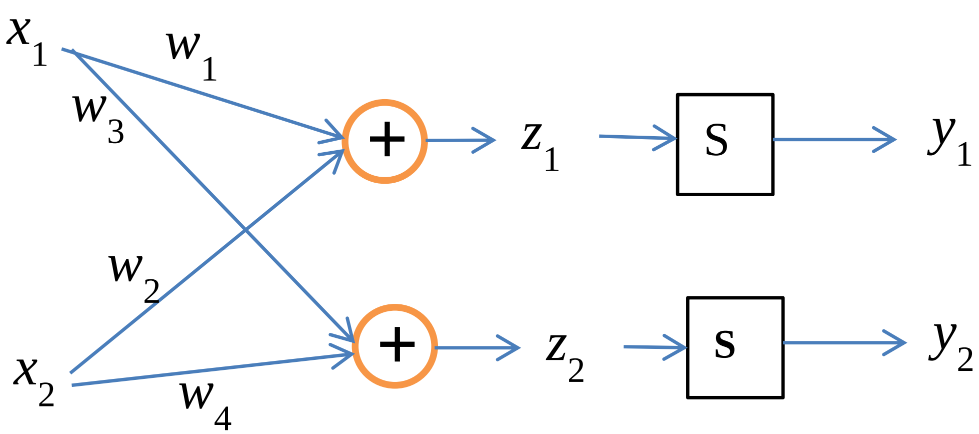
Shingchern D. You

Motivation

- The general form of back propagation is difficult to understand because of the notation
- We need to consider the following indices: Layer index, input index, output index, weights index, and iteration index
- Therefore, a notation like $w_{i,j}^{L}(k+1)$ might be used in the literature
- To avoid unnecessary confusion, we intend to make the notation simple and easy to follow

Forward computation

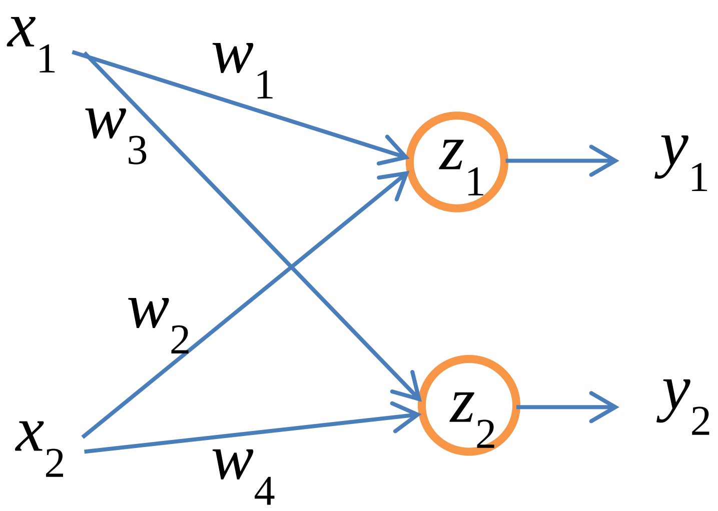
- The following is a simple example



- $z_1 = x_1 w_1 + x_2 w_2$
- $y_1 = \frac{1}{1+\exp(-z_1)}$

Forward computation

- Simplify the drawings



Single layer back propagation

- We use mean-square error as an example
- $\varepsilon = \varepsilon_1 + \varepsilon_2 = \frac{1}{2}((y_1 - d_1)^2 + (y_2 - d_2)^2)$
- We add $\frac{1}{2}$ to remove the constant in derivatives
- Want to minimize ε with respect to weights, we do gradient search $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)$
- In the present case, $f(\cdot) = \varepsilon(\cdot)$ and $\mathbf{x}_k$ is $\boldsymbol{w}_k$

Single layer back propagation

- To simplify the discussion, consider only updating w_1
- Therefore, $w_{1,(k+1)} \leftarrow w_{1,(k)} - \eta \frac{\partial}{\partial w_1} \varepsilon$
- We know $\frac{\partial}{\partial w_1} \varepsilon = \frac{\partial}{\partial w_1} \varepsilon_1$ because ε_2 is not related to w_1

Single layer back propagation

- Recall that we have

$$\varepsilon_1 = \frac{1}{2}(y_1 - d_1)^2$$

where d_1 is constant (desired output)

$$y_1 = \frac{1}{1 + \exp(-z_1)}$$

$$z_1 = x_1 w_1 + x_2 w_2$$

Single layer back propagation

- By chain rule, we have $\frac{\partial \varepsilon_1}{\partial w_1} = \frac{\partial \varepsilon_1}{\partial y_1}\frac{\partial y_1}{\partial z_1}\frac{\partial z_1}{\partial w_1}$

where

$$\frac{\partial \varepsilon_1}{\partial y_1} = (y_1 - d_1)$$

$$\frac{\partial y_1}{\partial z_1} = y_1(1 - y_1)$$

$$\frac{\partial z_1}{\partial w_1} = x_1$$

Single layer back propagation

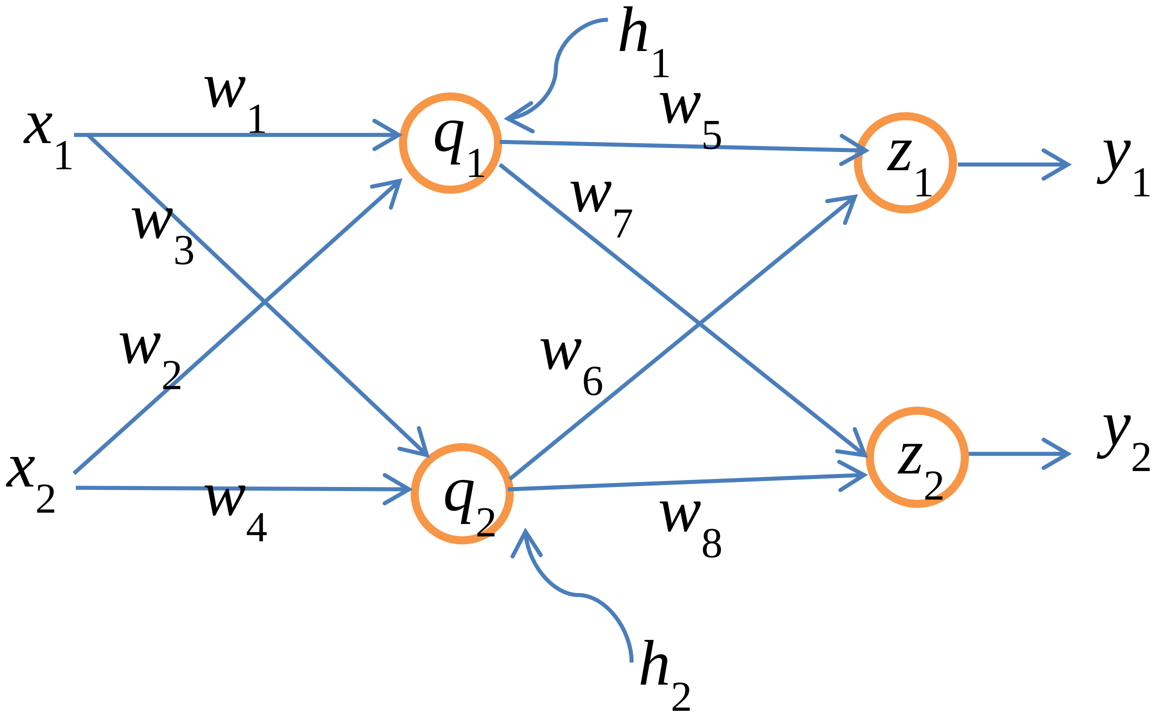
- Finally, we obtain

$$\frac{\partial \varepsilon_1}{\partial w_1} = (y_1 - d_1)y_1(1 - y_1)x_1$$

- For the k-th iteration, we should have all values of y_1, x_1, and d_1 at hands because y_1 is from forward computation and x_1 and d_1 are the input (instance) and desired output
- We can derived the update rule for other weights by the same method

Multi-layer back propagation

- We now extend the concept to multi-layer networks



Multi-layer back propagation

- What do we have now

$$q_1 = x_1 w_1 + x_2 w_2$$

$$h_1 = \frac{1}{1 + \exp(-q_1)}$$

$$z_1 = h_1 w_5 + h_2 w_6$$

$$y_1 = \frac{1}{1 + \exp(-z_1)}$$

$$\varepsilon_1 = \frac{1}{2}(y_1 - d_1)^2$$

Multi-layer back propagation

- From the single-layer results, we know

$$w_{5,(k+1)} \leftarrow w_{5,(k)} - \eta \frac{\partial \varepsilon}{\partial w_5}$$

where $\frac{\partial \varepsilon}{\partial w_5} = \frac{\partial \varepsilon_1}{\partial w_5} = \frac{\partial \varepsilon_1}{\partial y_1} \frac{\partial y_1}{\partial z_1} \frac{\partial z_1}{\partial w_5}$

$$= (y_1 - d_1) y_1 (1 - y_1) h_1$$

- Other weights in the second layer can be obtained by using the same approach

Multi-layer back propagation

- How about weights in the first (hidden) layer
- Use w_1 as an example: $\frac{\partial \varepsilon}{\partial w_1} = \frac{\partial \varepsilon_1}{\partial w_1} + \frac{\partial \varepsilon_2}{\partial w_1}$
- We know (again by chain rule)

$$\frac{\partial \varepsilon_1}{\partial w_1} = \frac{\partial \varepsilon_1}{\partial y_1} \frac{\partial y_1}{\partial z_1} \frac{\partial z_1}{\partial h_1} \frac{\partial h_1}{\partial q_1} \frac{\partial q_1}{\partial w_1}$$

and

$$\frac{\partial \varepsilon_2}{\partial w_1} = \frac{\partial \varepsilon_2}{\partial y_2} \frac{\partial y_2}{\partial z_2} \frac{\partial z_2}{\partial h_1} \frac{\partial h_1}{\partial q_1} \frac{\partial q_1}{\partial w_1}$$

Multi-layer back propagation

- Note that we can reuse partial results in weights updating in back propagation
- Observe the following equations

$$\frac{\partial \varepsilon_1}{\partial w_5} = \boldsymbol{\frac{\partial \varepsilon_1}{\partial y_1}\frac{\partial y_1}{\partial z_1}}\frac{\partial z_1}{\partial w_5}$$

$$\frac{\partial \varepsilon_1}{\partial w_1} = \boldsymbol{\frac{\partial \varepsilon_1}{\partial y_1}\frac{\partial y_1}{\partial z_1}}\frac{\partial z_1}{\partial h_1}\frac{\partial h_1}{\partial q_1}\frac{\partial q_1}{\partial w_1}$$

Multi-layer back propagation

- With the understanding of our example, you should be able to appreciate the "full" comprehensive BP equations given in the literature
- Notice that with more and more layers, the delta weight contains more and more terms, and thus, gets smaller and smaller
- That is one problem when training deep neural networks (i.e., networks with many layers)