

Deep Learning

06 Backpropagation-2

Dr. Konda Reddy Mopuri
Dept. of Artificial Intelligence
IIT Hyderabad
Jan-May 2025

Consider a specific Layer

- $x^{(l-1)} \xrightarrow{W^{(l)}, \mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$

Consider a specific Layer

- $x^{(l-1)} \xrightarrow{W^{(l)}, \mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $x_i^{(l)} = \sigma(s_i^{(l)})$

Consider a specific Layer

- $x^{(l-1)} \xrightarrow{W^{(l)}, \mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $x_i^{(l)} = \sigma(s_i^{(l)})$
- Since $s^{(l)}$ influences loss \mathcal{L} through only $x^{(l)}$,

$$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)})$$

Consider a specific Layer

- $x^{(l-1)} \xrightarrow{W^{(l)}, \mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $x_i^{(l)} = \sigma(s_i^{(l)})$
- Since $s^{(l)}$ influences loss \mathcal{L} through only $x^{(l)}$,

$$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)})$$

●

$$s_i^{(l)} = \sum_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}$$

Consider a specific Layer

- $x^{(l-1)} \xrightarrow{W^{(l)}, \mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $x_i^{(l)} = \sigma(s_i^{(l)})$
- Since $s^{(l)}$ influences loss \mathcal{L} through only $x^{(l)}$,

$$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)})$$

●

$$s_i^{(l)} = \sum_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}$$

- Since $x^{(l-1)}$ influences the loss \mathcal{L} only through $s^{(l)}$,

$$\frac{\partial \ell}{\partial x_j^{(l-1)}} = \sum_i \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial x_j^{(l-1)}} = \sum_i \frac{\partial \ell}{\partial s_i^{(l)}} W_{i,j}^{(l)}$$

We need gradients wrt parameters W and b

- $x^{(l-1)} \xrightarrow{W^{(l)}, b^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$

We need gradients wrt parameters W and b

- $x^{(l-1)} \xrightarrow{W^{(l)}, \mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $W_{i,j}^{(l)}$ and $\mathbf{b}^{(l)}$ influence the loss through $s^{(l)}$ via $s_i^{(l)} = \sum_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}$,

We need gradients wrt parameters W and b

- $x^{(l-1)} \xrightarrow{W^{(l)}, \mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $W_{i,j}^{(l)}$ and $\mathbf{b}^{(l)}$ influence the loss through $s^{(l)}$ via $s_i^{(l)} = \sum_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}$,

●

$$\frac{\partial \ell}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)} \quad (1)$$

We need gradients wrt parameters W and b



- $x^{(l-1)} \xrightarrow{W^{(l)}, \mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $W_{i,j}^{(l)}$ and $\mathbf{b}^{(l)}$ influence the loss through $s^{(l)}$ via $s_i^{(l)} = \sum_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}$,

- $$\frac{\partial \ell}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)} \quad (1)$$

- $$\frac{\partial \ell}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \quad (2)$$

Summary of Backprop

- From the definition of loss, obtain $\frac{\partial l}{\partial x_i^{(l)}}$

Summary of Backprop

- From the definition of loss, obtain $\frac{\partial \ell}{\partial x_i^{(l)}}$
- Recursively compute the loss derivatives wrt the activations

$$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)}) \text{ and } \frac{\partial \ell}{\partial x_j^{(l-1)}} = \sum_i \frac{\partial \ell}{\partial s_i^{(l)}} w_{i,j}^{(l)}$$

Summary of Backprop

- From the definition of loss, obtain $\frac{\partial \ell}{\partial x_i^{(l)}}$
- Recursively compute the loss derivatives wrt the activations

$$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)}) \text{ and } \frac{\partial \ell}{\partial x_j^{(l-1)}} = \sum_i \frac{\partial \ell}{\partial s_i^{(l)}} w_{i,j}^{(l)}$$

- Then wrt the parameters

$$\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)} \text{ and } \frac{\partial \ell}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}}$$

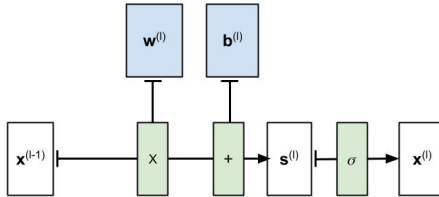
Jacobian in Tensorial form

- $\psi : \mathcal{R}^N \rightarrow \mathcal{R}^M$ then $\left[\frac{\partial \psi}{\partial x} \right] = \begin{bmatrix} \frac{\partial \psi_1}{\partial x_1} & \cdots & \frac{\partial \psi_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial \psi_M}{\partial x_1} & \cdots & \frac{\partial \psi_M}{\partial x_N} \end{bmatrix}$

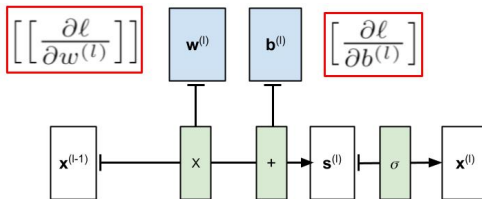
Jacobian in Tensorial form

- $\psi : \mathcal{R}^N \rightarrow \mathcal{R}^M$ then $\left[\frac{\partial \psi}{\partial x} \right] = \begin{bmatrix} \frac{\partial \psi_1}{\partial x_1} & \cdots & \frac{\partial \psi_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial \psi_M}{\partial x_1} & \cdots & \frac{\partial \psi_M}{\partial x_N} \end{bmatrix}$
- $\psi : \mathcal{R}^{N \times M} \rightarrow \mathcal{R}$ then $\left[\left[\frac{\partial \psi}{\partial x} \right] \right] = \begin{bmatrix} \frac{\partial \psi}{\partial w_{1,1}} & \cdots & \frac{\partial \psi}{\partial w_{1,M}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \psi}{\partial w_{N,1}} & \cdots & \frac{\partial \psi}{\partial w_{N,M}} \end{bmatrix}$

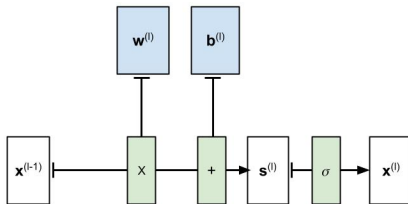
Forward Pass



Goal of Backward Pass

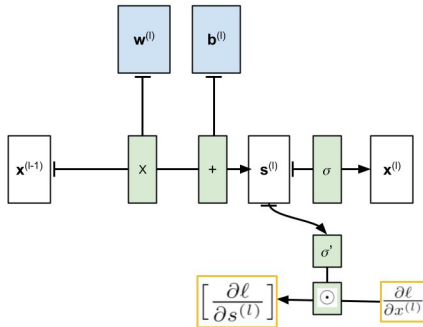


Begin from succeeding layer

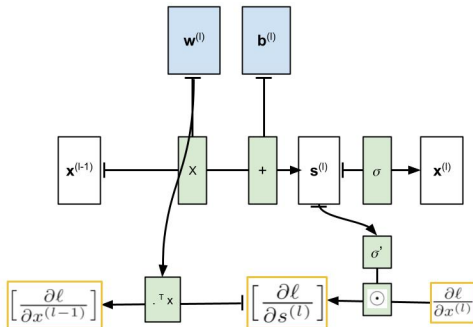


$$\frac{\partial \ell}{\partial \mathbf{x}^{(l)}}$$

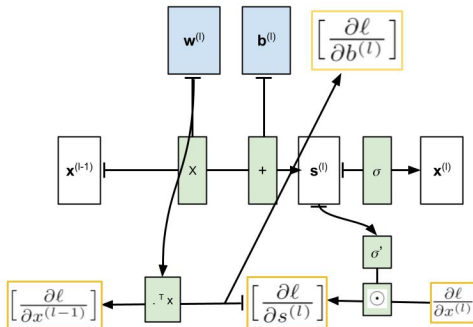
Begin from succeeding layer



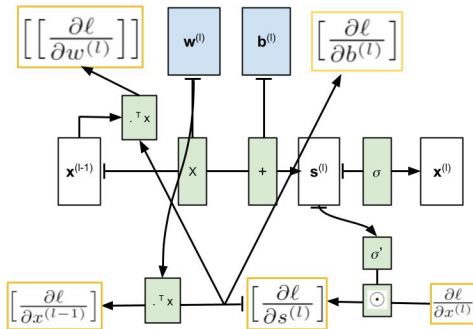
Begin from succeeding layer



Begin from succeeding layer



Begin from succeeding layer



Update the parameters

- $W^{(l)} = W^{(l)} - \eta \left[\left[\frac{\partial \ell}{\partial w^{(l)}} \right] \right]$ and $\mathbf{b}^{(l)} = \mathbf{b}^{(l)} - \eta \left[\frac{\partial \ell}{\partial b^{(l)}} \right]$

Observations

- BP is basically simple: applying chain rule iteratively

- BP is basically simple: applying chain rule iteratively
- It can be expressed in tensorial form (similar to the forward pass)

- BP is basically simple: applying chain rule iteratively
- It can be expressed in tensorial form (similar to the forward pass)
- Heavy computations are with the linear operations

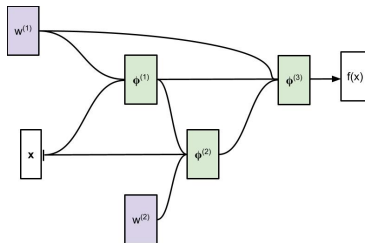
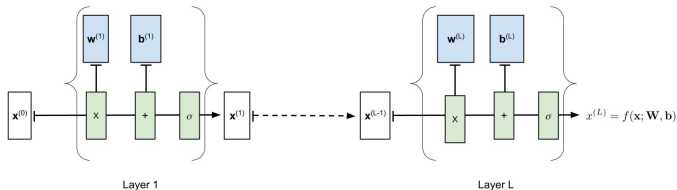
- BP is basically simple: applying chain rule iteratively
- It can be expressed in tensorial form (similar to the forward pass)
- Heavy computations are with the linear operations
- Nonlinearities go into simple element wise operations

- BP is basically simple: applying chain rule iteratively
- It can be expressed in tensorial form (similar to the forward pass)
- Heavy computations are with the linear operations
- Nonlinearities go into simple element wise operations
- BP Needs all the intermediate layer results to be in memory

- BP is basically simple: applying chain rule iteratively
- It can be expressed in tensorial form (similar to the forward pass)
- Heavy computations are with the linear operations
- Nonlinearities go into simple element wise operations
- BP Needs all the intermediate layer results to be in memory
- Takes twice the computations of forward pass

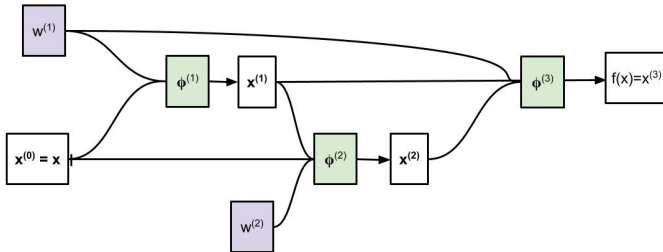
Beyond MLP

- We can generalize MLP



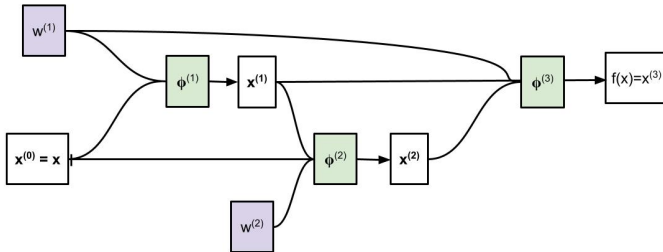
To an arbitrary Directed Acyclic Graph (DAG)

Forward pass in the computational graph



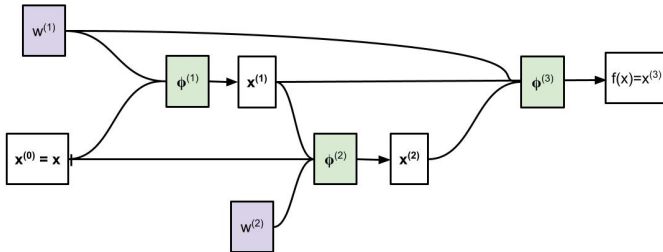
● $x^{(0)} = x$

Forward pass in the computational graph



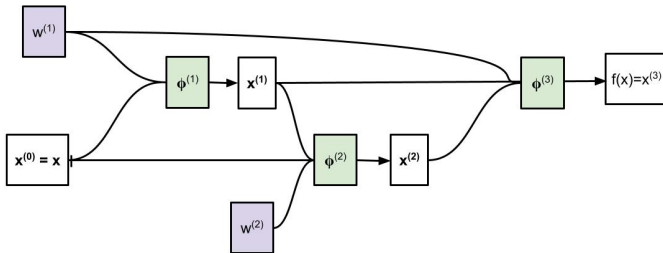
- $x^{(0)} = x$
- $x^{(1)} = \phi^{(1)}(x^{(0)}; w^{(1)})$

Forward pass in the computational graph



- $x^{(0)} = x$
- $x^{(1)} = \phi^{(1)}(x^{(0)}; w^{(1)})$
- $x^{(2)} = \phi^{(2)}(x^{(0)}, x^{(1)}; w^{(2)})$

Forward pass in the computational graph



- $x^{(0)} = x$
- $x^{(1)} = \phi^{(1)}(x^{(0)}; w^{(1)})$
- $x^{(2)} = \phi^{(2)}(x^{(0)}, x^{(1)}; w^{(2)})$
- $f(x) = x^{(3)} = \phi^{(3)}(x^{(1)}, x^{(2)}; w^{(1)})$

Notation: Jacobian of a general transformation



if $(a_1 \dots a_Q) = \phi(b_1 \dots b_R)$ then we use the notation (3)

$$\left[\frac{\partial a}{\partial b} \right] = J_\phi^T = \begin{bmatrix} \frac{\partial a_1}{\partial b_1} & \dots & \frac{\partial a_Q}{\partial b_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial a_1}{\partial b_R} & \dots & \frac{\partial a_Q}{\partial b_R} \end{bmatrix} \quad (4)$$

Notation: Jacobian of a general transformation

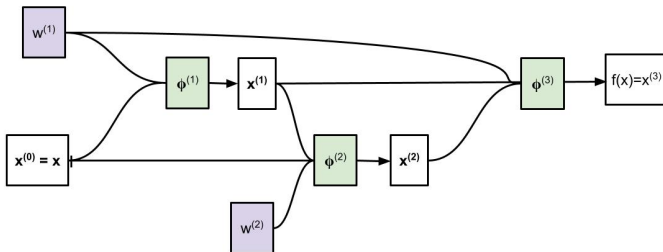
if $(a_1 \dots a_Q) = \phi(b_1 \dots b_R)$ then we use the notation (3)

$$\left[\frac{\partial a}{\partial b} \right] = J_{\phi}^T = \begin{bmatrix} \frac{\partial a_1}{\partial b_1} & \dots & \frac{\partial a_Q}{\partial b_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial a_1}{\partial b_R} & \dots & \frac{\partial a_Q}{\partial b_R} \end{bmatrix} \quad (4)$$

if $(a_1 \dots a_Q) = \phi(b_1 \dots b_R; c_1 \dots c_S)$ then we use the notation (5)

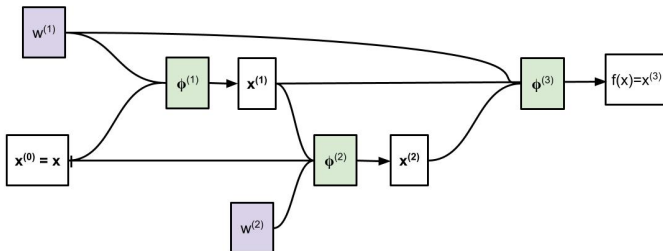
$$\left[\frac{\partial a}{\partial c} \right] = J_{\phi|c}^T = \begin{bmatrix} \frac{\partial a_1}{\partial c_1} & \dots & \frac{\partial a_Q}{\partial c_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial a_1}{\partial c_S} & \dots & \frac{\partial a_Q}{\partial c_S} \end{bmatrix} \quad (6)$$

Backward pass



- From the loss equation, we can compute $\left[\frac{\partial \ell}{\partial x^{(3)}} \right]$

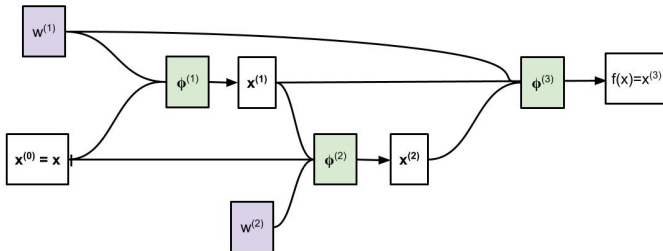
Backward pass



- From the loss equation, we can compute $\left[\frac{\partial \ell}{\partial x^{(3)}} \right]$

$$\left[\frac{\partial \ell}{\partial x^{(2)}} \right] = \left[\frac{\partial x^{(3)}}{\partial x^{(2)}} \right] \left[\frac{\partial \ell}{\partial x^{(3)}} \right] = J_{\phi^{(3)}|x^{(2)}}^T \left[\frac{\partial \ell}{\partial x^{(3)}} \right]$$

Backward pass

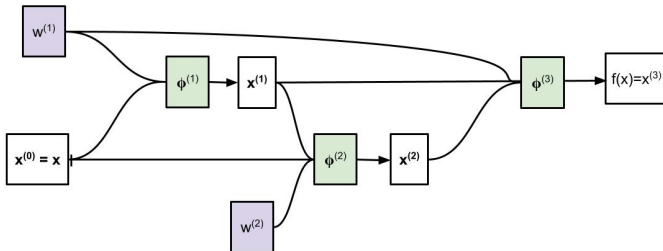


- From the loss equation, we can compute $\left[\frac{\partial \ell}{\partial x^{(3)}} \right]$

$$\left[\frac{\partial \ell}{\partial x^{(2)}} \right] = \left[\frac{\partial x^{(3)}}{\partial x^{(2)}} \right] \left[\frac{\partial \ell}{\partial x^{(3)}} \right] = J_{\phi^{(3)}|x^{(2)}}^T \left[\frac{\partial \ell}{\partial x^{(3)}} \right]$$

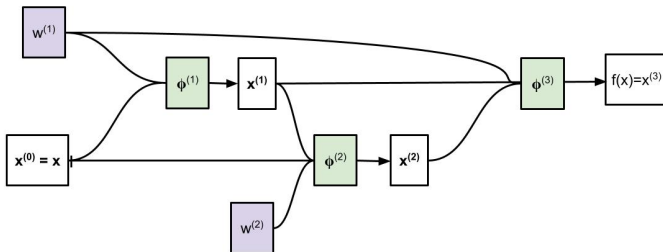
$$\begin{aligned} \left[\frac{\partial \ell}{\partial x^{(1)}} \right] &= \left[\frac{\partial x^{(3)}}{\partial x^{(1)}} \right] \left[\frac{\partial \ell}{\partial x^{(3)}} \right] + \left[\frac{\partial x^{(2)}}{\partial x^{(1)}} \right] \left[\frac{\partial \ell}{\partial x^{(2)}} \right] \\ &= J_{\phi^{(3)}|x^{(1)}}^T \left[\frac{\partial \ell}{\partial x^{(3)}} \right] + J_{\phi^{(2)}|x^{(1)}}^T \left[\frac{\partial \ell}{\partial x^{(2)}} \right] \end{aligned}$$

Backward pass



$$\begin{aligned} \left[\frac{\partial \ell}{\partial w^{(1)}} \right] &= \left[\frac{\partial x^{(3)}}{\partial w^{(1)}} \right] \left[\frac{\partial \ell}{\partial x^{(3)}} \right] + \left[\frac{\partial x^{(1)}}{\partial w^{(1)}} \right] \left[\frac{\partial \ell}{\partial x^{(1)}} \right] \\ &= J_{\phi^{(3)}|w^{(1)}}^T \left[\frac{\partial \ell}{\partial x^{(3)}} \right] + J_{\phi^{(1)}|w^{(1)}}^T \left[\frac{\partial \ell}{\partial x^{(1)}} \right] \end{aligned}$$

Backward pass



$$\begin{aligned} \left[\frac{\partial \ell}{\partial w^{(1)}} \right] &= \left[\frac{\partial x^{(3)}}{\partial w^{(1)}} \right] \left[\frac{\partial \ell}{\partial x^{(3)}} \right] + \left[\frac{\partial x^{(1)}}{\partial w^{(1)}} \right] \left[\frac{\partial \ell}{\partial x^{(1)}} \right] \\ &= J_{\phi^{(3)}|w^{(1)}}^T \left[\frac{\partial \ell}{\partial x^{(3)}} \right] + J_{\phi^{(1)}|w^{(1)}}^T \left[\frac{\partial \ell}{\partial x^{(1)}} \right] \end{aligned}$$

$$\left[\frac{\partial \ell}{\partial w^{(2)}} \right] = \left[\frac{\partial x^{(2)}}{\partial w^{(2)}} \right] \left[\frac{\partial \ell}{\partial x^{(2)}} \right] = J_{\phi^{(2)}|w^{(2)}}^T \left[\frac{\partial \ell}{\partial x^{(2)}} \right]$$

Observations, some more

- Does BP always find the 'right' function? (Let's assume it converged to the global minimum of the loss function)

Observations, some more

- Does BP always find the 'right' function? (Let's assume it converged to the global minimum of the loss function)
- Remember, our loss function is only a proxy for the classification error

Observations, some more

- Does BP always find the 'right' function? (Let's assume it converged to the global minimum of the loss function)
- Remember, our loss function is only a proxy for the classification error
- Minimizing the proxy may not minimize the actual

Observations, some more

- Does BP always find the 'right' function? (Let's assume it converged to the global minimum of the loss function)
- Remember, our loss function is only a proxy for the classification error
- Minimizing the proxy may not minimize the actual
- i.e., ideal function (separation for classification) may not be a feasible optimum for the chosen loss function

Observations, some more

- New training samples may change BP minimally

Observations, some more

- New training samples may change BP minimally
- Prefers consistency (low variance) over perfection (low bias)

Observations, some more

- New training samples may change BP minimally
- Prefers consistency (low variance) over perfection (low bias)
- Minimizing the proxy may not minimize the actual

High dimensional loss surfaces are complex

- Saddle points are far more frequent than local minima (exponential in network size)

High dimensional loss surfaces are complex

- Saddle points are far more frequent than local minima (exponential in network size)
- Most local minima are equivalent and lie close to the global minimum

High dimensional loss surfaces are complex

- Saddle points are far more frequent than local minima (exponential in network size)
- Most local minima are equivalent and lie close to the global minimum
- Active research!