Deep Learning

3.4 Backpropagation

Dr. Konda Reddy Mopuri
kmopuri@iittp.ac.in
Dept. of CSE, IIT Tirupati

Recap

1. Gradient of a scalar valued function $f(\mathbf{x})$: $\mathbf{x} \rightarrow \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_D}\right)$

Recap

1. Gradient of a scalar valued function $f(\mathbf{x})$: $\mathbf{x} \rightarrow \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_D}\right)$
2. Gradient of a vector valued function $\mathbf{f}(\mathbf{x})$ is called Jacobian:

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^{\mathrm{T}} f_1 \\ \vdots \\ \nabla^{\mathrm{T}} f_m \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Gradient descent on MLP

1. Loss is $\mathcal{L}(W, \mathbf{b}) = \sum_n l(f(x_n; W, \mathbf{b}), y_n)$

Gradient descent on MLP

1. Loss is $\mathcal{L}(W, \mathbf{b}) = \sum_n l(f(x_n; W, \mathbf{b}), y_n)$
2. For applying Gradient descent, we need gradient of individual sample loss with respect to all the model parameters

$$l_n = l(f(x_n; W, \mathbf{b}), y_n)$$

$\frac{\partial l_n}{\partial W_{i,j}^{(l)}}$ and $\frac{\partial l_n}{\partial \mathbf{b}_i^{(l)}}$

Forward pass operation

$$x^{(0)} = x \xrightarrow{W^{(1)},\mathbf{b}^{(1)}} s^{(1)} \xrightarrow{\sigma} x^{(1)} \xrightarrow{W^{(2)},\mathbf{b}^{(2)}} s^{(2)} \dots x^{(L-1)} \xrightarrow{W^{(L)},\mathbf{b}^{(L)}} s^{(L)} \xrightarrow{\sigma} x^{(L)} = f(x;W,\mathbf{b})$$

Formally, $x^{(0)} = x, f(x;W,\mathbf{b}) = x^{(L)}$

$$\forall l = 1,\dots,L \quad \begin{cases} s^{(l)} & = W^{(l)}x^{(l-1)} + \mathbf{b}^{(l)} \\ x^{(l)} & = \sigma(s^{(l)}) \end{cases}$$

Chain rule of differential calculus

1. Core concept of backpropagation

Chain rule of differential calculus

1. Core concept of backpropagation
2. $$(g \circ f)'(x) = g'(f(x)) \cdot f'(x)$$

Chain rule of differential calculus

1. Core concept of backpropagation
2. $$(g \circ f)'(x) = g'(f(x)) \cdot f'(x)$$
3. $$\frac{\partial}{\partial x} g(f(x)) = \left.\frac{\partial g(a)}{\partial a}\right|_{a=f(x)} \cdot \frac{\partial f(x)}{\partial x}$$

Chain rule of differential calculus

1. Core concept of backpropagation
2. $$(g \circ f)'(x) = g'(f(x)) \cdot f'(x)$$
3. $$\frac{\partial}{\partial x} g(f(x)) = \left.\frac{\partial g(a)}{\partial a}\right|_{a=f(x)} \cdot \frac{\partial f(x)}{\partial x}$$
4. $$J_{f_N \circ f_{N-1} \circ \ldots f_1}(x) = J_{f_N}(f_{N-1}(\ldots(x)))\ldots J_{f_3}(f_2(f_1(x))) \cdot J_{f_2}(f_1(x)) \cdot J_{f_1}(x)$$

Chain rule of differential calculus

1. Core concept of backpropagation
2. $$(g \circ f)'(x) = g'(f(x)) \cdot f'(x)$$
3. $$\frac{\partial}{\partial x} g(f(x)) = \left.\frac{\partial g(a)}{\partial a}\right|_{a=f(x)} \cdot \frac{\partial f(x)}{\partial x}$$
4. $$J_{f_N \circ f_{N-1} \circ \ldots f_1}(x) = J_{f_N}(f_{N-1}(\ldots(x)))\ldots J_{f_3}(f_2(f_1(x))) \cdot J_{f_2}(f_1(x)) \cdot J_{f_1}(x)$$

- $J_{f(x)}$ is Jacobian of f computed at x.

Consider a specific Layer

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

Consider a specific Layer

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

Consider a specific Layer

1. $x^{(l-1)} \xrightarrow{W^{(l)},\mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
2. $x_i^{(l)} = \sigma(s_i^{(l)})$
3. 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

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

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

4.

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

Consider a specific Layer

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

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

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

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

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

We need gradients wrt parameters W and b

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

We need gradients wrt parameters W and b

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

We need gradients wrt parameters W and b

1. $x^{(l-1)} \xrightarrow{W^{(l)},\mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
2. $W_{i,j}^{(l)}$ and $\mathbf{b}^{(l)}$ influence the loss through $s^{(l)}$ via $s_i^{(l)} = \Sigma_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}$,
3. $$\frac{\partial \ell}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)} \tag{1}$$

We need gradients wrt parameters W and b

1. $x^{(l-1)} \xrightarrow{W^{(l)},\mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
2. $W_{i,j}^{(l)}$ and $\mathbf{b}^{(l)}$ influence the loss through $s^{(l)}$ via $s_i^{(l)} = \Sigma_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}$,
3. $$\frac{\partial \ell}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)} \tag{1}$$
4. $$\frac{\partial \ell}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \tag{2}$$

Summary of Backprop

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

Summary of Backprop

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

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

Summary of Backprop

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

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

3. Then wrt the parameters
$$\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)} \text{ and } \frac{\partial \ell}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}}$$

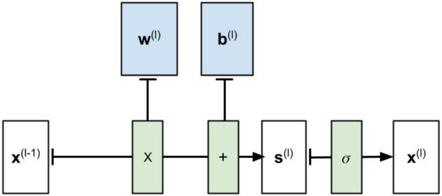
Jocobian in Tensorial form

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

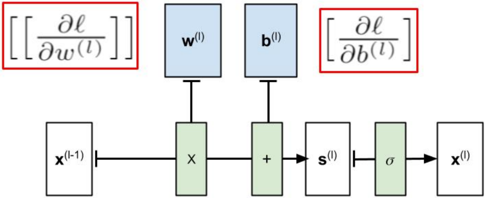
Jocobian in Tensorial form

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

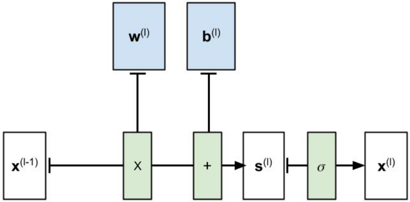
Forward Pass



Goal of Backward Pass

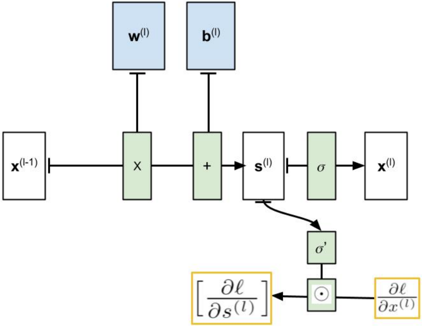


Begin from succeeding layer

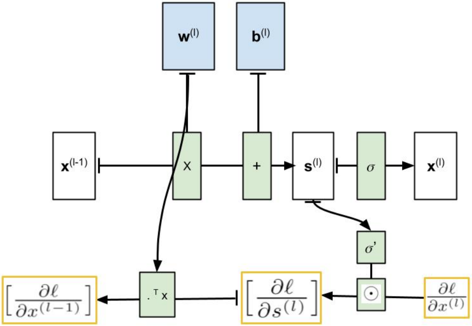


$$\boxed{\frac{\partial \ell}{\partial x^{(l)}}}$$

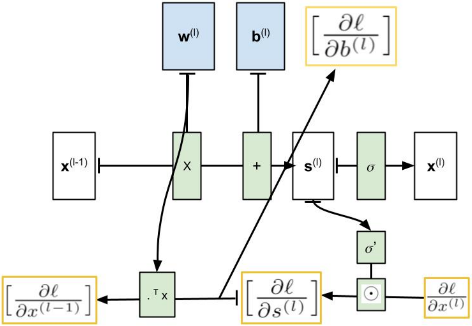
Begin from succeeding layer



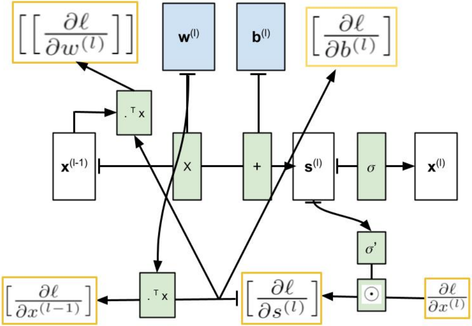
Begin from succeeding layer



Begin from succeeding layer



Begin from succeeding layer



Update the parameters

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

Observations

1. BP is basically simple: applying chain rule iteratively

Observations

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

Observations

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

Observations

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

Observations

1. BP is basically simple: applying chain rule iteratively
2. It can be expressed in tensorial form (similar to the forward pass)
3. Heavy computations are with the linear operations
4. Nonlinearities go into simple element wise operations
5. In an untreated situation, BP Needs all the intermediate layer results to be in memory

Observations

1. BP is basically simple: applying chain rule iteratively
2. It can be expressed in tensorial form (similar to the forward pass)
3. Heavy computations are with the linear operations
4. Nonlinearities go into simple element wise operations
5. In an untreated situation, BP Needs all the intermediate layer results to be in memory
6. Takes twice the computations of forward pass