Section B: Backprop

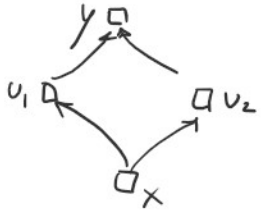
Sunday, October 9, 2022 8:44 PM

Chain Rule

Def #1: $y = f(u)$, $u = g(x)$

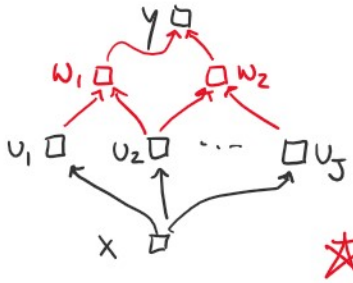
$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

Def #2: $y = f(u_1, u_2)$, $u_1 = g_1(x)$, $u_2 = g_2(x)$



$$\frac{dy}{dx} = \frac{dy}{du_1}\frac{du_1}{dx} + \frac{dy}{du_2}\frac{du_2}{dx}$$

Def #3: $y = f(\vec{u})$, $u = g(x)$



$$\frac{dy}{dx} = \sum_{j=1}^{J} \frac{dy}{du_j}\frac{du_j}{dx}$$

★ holds for any such intermediate vars. w_j

Backprop Ex #1

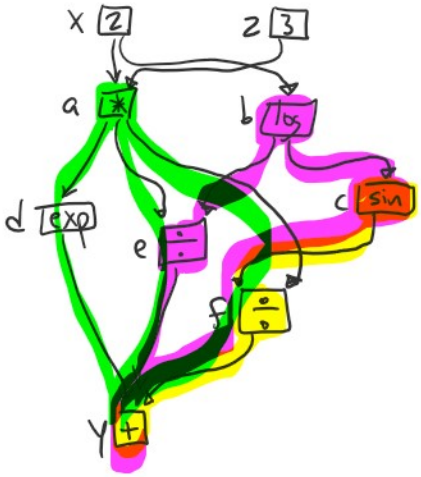
$$y = f(x,z) = \exp(xz) + \frac{xz}{\log(x)} + \frac{\sin(\log(x))}{xz}$$

Forward Computation

Given $x=2, z=3$

- ✓ $a = xz$
- ✓ $b = \log(x)$
- ✓ $c = \sin(b)$
- ✓ $d = \exp(a)$
- ✓ $e = a/b$
- ✓ $f = c/a$
- ✓ $y = d+e+f$

Computation Graph



Backward Computation

$$g_y = \frac{dy}{dy} = 1$$

$g_f = \frac{dy}{df} = 1$, $g_e = \frac{dy}{de} = 1$, $g_d = \frac{dy}{dd} = 1$

$$g_c = \frac{dy}{dc} = \frac{dy}{df}\frac{df}{dc} = (g_f)\left(\frac{1}{a}\right)$$

$$g_b = \frac{dy}{db} = \frac{dy}{de}\frac{de}{db} + \frac{dy}{dc}\frac{dc}{db} = (g_e)\left(-\frac{a}{b^2}\right) + (g_c)(\cos(b))$$

$$g_a = \frac{dy}{da} = \frac{dy}{dd}\frac{dd}{da} + \frac{dy}{de}\frac{de}{da} + \frac{dy}{df}\frac{df}{da}$$

Updates for Backprop

Updates for Backprop

$$g_x = \frac{dy}{dx} = \sum_{k=1}^{K} \frac{dy}{du_k} \frac{du_k}{dx}$$

$$\bigstar = \sum_{k=1}^{K} (g_{u_k}) \left(\frac{du_k}{dx}\right)$$

Efficient b/c
- reuse of forward comp.
- reuse of backward comp.

$$g_a = \frac{dy}{da} = \frac{dy}{dd}\frac{dd}{da} + \frac{dy}{de}\frac{de}{da} + \frac{dy}{df}\frac{df}{da}$$

$$= (g_d)(\exp(a)) + (g_e)\left(\frac{1}{b}\right) + (g_f)\left(-\frac{c}{a^2}\right)$$

$$g_x = (g_a)(z) + (g_b)\left(\frac{1}{x}\right)$$

$$g_z = (g_a)(x)$$

Neural Network Training

- Consider a 2-hidden layer NN
- params are $\theta = [\alpha^{(1)}, \alpha^{(2)}, \beta]$
- SGD Training:

  Iterate until convergence:

  1. Sample $i \in \{1, \ldots, N\}$
  2. Compute gradient by backprop:

$$g_{\alpha^{(1)}} = \nabla_{\alpha^{(1)}} J^{(i)}(\theta)$$

$$g_{\alpha^{(2)}} = \nabla_{\alpha^{(2)}} J^{(i)}(\theta)$$

$$g_\beta = \nabla_\beta J^{(i)}(\theta)$$

$$J^{(i)}(\theta) = \ell\left(h_\theta(\vec{x}^{(i)}), y^{(i)}\right)$$

  3. Update parameters

$$\alpha^{(1)} \leftarrow \alpha^{(1)} - \gamma g_{\alpha^{(1)}}$$

$$\alpha^{(2)} \leftarrow \alpha^{(2)} - \gamma g_{\alpha^{(2)}}$$

$$\beta \leftarrow \beta - \gamma g_\beta$$

Aside:

$$\nabla J(\vec{a}, \vec{b}) = \nabla_{\vec{a},\vec{b}} J(\vec{a}, \vec{b})$$

$$\nabla_{\vec{a}} J(\vec{a}, \vec{b}) = \begin{bmatrix} \partial J(\vec{a},\vec{b})/\partial a_1 \\ \vdots \\ \partial J(\vec{a},\vec{b})/\partial a_K \end{bmatrix}$$

Backprop Ex#2 for NN

Given:
1. Dec. fn. $\hat{y} = h_\theta(\vec{x}) = \sigma\left((\alpha^{(3)})^T \sigma\left((\alpha^{(2)})^T \sigma\left((\alpha^{(1)})^T \vec{x}\right)\right)\right)$
2. Loss. fn. $J = \ell(\hat{y}, y^*) = -(y^* \log(\hat{y}) + (1-y^*)\log(1-\hat{y}))$
3. Training ex. $(\vec{x}, y^*)$

left out intercept terms

$\vec{\beta}$ $\vec{z}^{(2)}$ $\vec{z}^{(1)}$

Loss: $J = \ell(\hat{y}, y^*) = -\left(y^* \log(\hat{y}) + (1-y^*)\log(1-\hat{y})\right)$

③ Training ex. $(\vec{x}, y^*)$

Forward Comp.

Given $\vec{x}, y^*, \alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}$

$z^{(0)} = \vec{x}$

for $i = 1, 2, 3$:

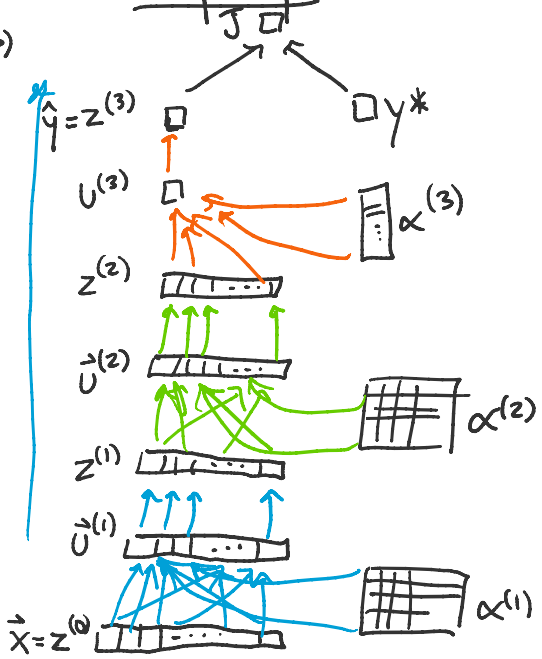
$\vec{u}^{(i)} = (\alpha^{(i)})^T z^{(i-1)}$

$z^{(i)} = \sigma(\vec{u}^{(i)})$

$\hat{y} = z^{(3)}$

$J = \ell(\hat{y}, y^*)$

Comp. Graph



Backward Comp.

$g_J = [1]$

$g_{\hat{y}} = -\left(\frac{y^*}{\hat{y}} + \frac{(1-y^*)}{(1-\hat{y})}\right)$

for $i = 3, 2, 1$:

$g_{\vec{u}^{(i)}} = \ldots$

$g_{z^{(i-1)}} = \ldots$

$g_{\alpha^{(i)}} = \ldots$

$g_x = g_{z^{(0)}}$

HWS

Vector Chain Rule

$$\underbrace{\frac{\partial y}{\partial \vec{x}}}_{P \times 1} = \underbrace{\left(\underbrace{\left(\frac{\partial y}{\partial \vec{u}}\right)^T}_{N \times 1} \underbrace{\left(\frac{\partial \vec{u}}{\partial \vec{x}}\right)^T}_{P \times N}\right)^T}_{1 \times P}$$

$P \times 1$

$$= \frac{\partial \vec{u}}{\partial \vec{x}} \frac{\partial y}{\partial \vec{u}}$$