Overview

- Backpropagation

Differential circuits / networks

composition of arithmetic operations and elementary functions

$\Downarrow$

$(+, -, \times, /)$

(cos, sin, exp, log, relu, sigmoid)

Theorem (informally stated)

Suppose a differentiable circuit of size N,
computes a real-valued function $f: \mathbb{R}^l \to \mathbb{R}$
Then the gradient of f can also be computed in time $O(N)$
by a circuit of size $O(N)$

- $f \to J^{(i)}(\theta)$ $\quad l$ = # of parameters
  $N \approx O($# of parameters$)$

  $\nabla J^{(i)}(\theta)$ $\quad$ time $\approx O(N)$

Preliminary: Chain rule

Suppose J is a function of $\theta_1, \dots, \theta_p$

intermediate $g_j = g_j(\theta_1, \dots, \theta_p) \quad j = 1, \dots, k$

output var $J = J(g_1, \dots, g_k)$

$$\frac{\partial J}{\partial \theta_i} = \sum_{j=1}^{k} \frac{\partial J}{\partial g_j} \cdot \frac{\partial g_j}{\partial \theta_i}$$

$\in \mathbb{R}$ $\quad \in \mathbb{R}$ $\quad \in \mathbb{R}$

One-neuron neuron network example

input: x, label: y

$z = w^T x + b$

$h_\theta(x) \to o = \text{relu}(z)$

$w \in \mathbb{R}^d$, $x \in \mathbb{R}^d$, $b \in \mathbb{R}$, $z \in \mathbb{R}$

Loss: $J(\theta) = \frac{1}{2}(y-o)^2$

o as intermediate variable

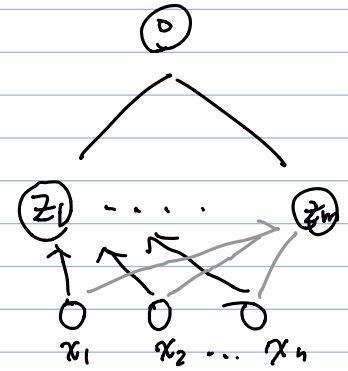
$$\frac{\partial J}{\partial w_i} = \frac{\partial J}{\partial o}\frac{\partial o}{\partial w_i} = (o-y)\frac{\partial o}{\partial w_i}$$

$$= (o-y)\frac{\partial o}{\partial z}\cdot\frac{\partial z}{\partial w_i}$$

$$= (o-y)\,\text{relu}'(z)\cdot x_i$$

$$= (o-y)\,\mathbb{1}(z\geq 0)\cdot x_i$$

$$\frac{\partial J}{\partial b} = \frac{\partial J}{\partial o}\frac{\partial o}{\partial b} = \frac{\partial J}{\partial o}\cdot\frac{\partial o}{\partial z}\cdot\frac{\partial z}{\partial b}$$



Two layer neural network

$w_j^{[1]} \in \mathbb{R}^d$

$b_j^{[1]} \in \mathbb{R}$

$$z_j = w_j^{[1]T}x + b_j^{[1]}$$

$$a_j = \text{relu}(z_j)$$

$$a = [a_1, \dots a_m]^T \in \mathbb{R}^m$$

$$o = w^{[2]T}a + b^{[2]}$$

$$J = \frac{1}{2}(y-o)^2$$

$$w_j^{[1]} = \left((w_j^{[1]})_1, \dots (w_j^{[1]})_d\right)^T \in \mathbb{R}^d$$

$$\frac{\partial J}{\partial (w_j^{[1]})_\ell} = \sum_{k=1}^{m}\frac{\partial J}{\partial z_k}\frac{\partial z_k}{\partial (w_j^{[1]})_\ell}$$

z_k and w_j has No dependency if $k \neq j$

$$= \frac{\partial J}{\partial z_j}\frac{\partial z_j}{\partial (w_j^{[1]})_\ell}$$

$$= \frac{\partial J}{\partial z_j}x_\ell$$

$$= \frac{\partial J}{\partial a_j}\frac{\partial a_j}{\partial z_j}\cdot x_\ell$$

$$= \frac{\partial J}{\partial a_j}\,\text{relu}'(z_j)\cdot x_\ell$$

$$= \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial a_j} \, \text{relu}'(z_j) \, x_l$$

$$= \boxed{(o-y)} \, (w^{[2]})_j \, \text{relu}'(z_j) \cdot x_l$$

Computing $\frac{\partial J}{\partial (w_j^{[1]})_l}$ takes O(# parameters) time

full gradient $\Rightarrow$ $O(\#\text{ parameters}^2)$ time.

$\forall j, l$ $(o-y)$ computation is shared

$\forall l$, fixed j $\text{relu}(z_j)$ shared $\left(\frac{\partial J}{\partial z_1}, \cdots \frac{\partial J}{\partial z_m}\right)$

For backprop two layer neural network with vectorized notations

$$\boxed{\delta^{[2]}} \triangleq \frac{\partial J}{\partial o} = (o-y) \in \mathbb{R}$$

Note $A \odot B = \begin{bmatrix} A_1 B_1 \\ \vdots \\ A_n B_n \end{bmatrix}$

$$\boxed{\delta^{[1]}} \triangleq \begin{bmatrix} \frac{\partial J}{\partial z_1} \\ \vdots \\ \frac{\partial J}{\partial z_m} \end{bmatrix} \in \mathbb{R}^m$$

$$\frac{\partial J}{\partial z} = \begin{bmatrix} \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial a_j} \cdot \frac{\partial a_j}{\partial z_j} \\ \vdots \end{bmatrix}$$

$$= (o-y)(w^{[2]T}) \odot \text{relu}(z)$$

① Compute $\frac{\partial J}{\partial o} = (o-y)$

② Compute $\begin{bmatrix} \frac{\partial J}{\partial z_1} \\ \vdots \\ \frac{\partial J}{\partial z_m} \end{bmatrix}$ $\frac{\partial J}{\partial z_j} = \boxed{\frac{\partial J}{\partial o}} \cdot \underbrace{\frac{\partial o}{\partial a_j}}_{w_j^{[2]}} \cdot \underbrace{\frac{\partial a_j}{\partial z_j}}_{\text{relu}'(z_j)} = (o-y) w_j^{[2]} \, \text{relu}'(z_j)$

③ $\boxed{\frac{\partial J}{\partial (w_j^{[1]})_l}} = \boxed{\frac{\partial J}{\partial z_j}} \cdot \underline{\frac{\partial z_j}{\partial (w_j^{[1]})_l}}$

$= \frac{\partial J}{\partial z_j} \cdot \boxed{x_l}$

$$\frac{\partial J}{\partial w_j^{[1]}} = \begin{matrix} \\ m \end{matrix} \overbrace{\begin{bmatrix} \frac{\partial J}{\partial z_j} \cdot \frac{\partial z}{\partial (w_j^{[1]})_1} \\ \vdots \end{bmatrix}}^{d} = \frac{\partial J}{\partial z} \cdot x^T$$

$$\frac{\partial J}{\partial b_j^{[1]}} = \boxed{\frac{\partial J}{\partial z_j}} \cdot \underbrace{\frac{\partial z_j}{\partial b_j}}_{\boxed{1}} \quad \left(z_j = w^{[1]} x + b_j^{[1]}\right)$$

$$\frac{\partial J}{\partial (w^{[2]})_j} = \quad \frac{\partial J}{\partial (b^{[2]})_j}$$

Some notations

$\frac{\partial A}{\partial B}$ A is a real value variable

B is a vector / matrix

$B \in \mathbb{R}^m \quad \frac{\partial A}{\partial B} = \begin{bmatrix} \frac{\partial A}{\partial B_1} \\ \vdots \\ \frac{\partial A}{\partial B_m} \end{bmatrix} \in \mathbb{R}^m$

$B \in \mathbb{R}^{m \times n} \quad \frac{\partial A}{\partial B} = \left[\frac{\partial A}{\partial B_{ij}} \right] \in \mathbb{R}^{m \times n}$

Vectorized Gradients

$\delta^{[2]} = o - y \quad \left(\frac{\partial J}{\partial o}\right)$

$\delta^{[1]} = (o-y)(w^{[2]})^T \odot \text{relu}'(z) \in \mathbb{R}^m \quad \left(\frac{\partial J}{\partial z}\right)$

$\frac{\partial J}{\partial w^{[1]}} = \frac{\partial J}{\partial z} \cdot x^T = \delta^{[1]} x^T \in \mathbb{R}^{m \times d}$

$\mathbb{R}^m \quad \mathbb{R}^{1 \times d}$

$\frac{\partial J}{\partial b^{[1]}} = \frac{\partial J}{\partial z} \cdot \frac{\partial z}{\partial b^{[1]}} = \delta^{[1]} \in \mathbb{R}$

$\frac{\partial J}{\partial w^{[2]}} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial w^{[2]}} = (o-y) \cdot a = \delta^{[2]} \cdot a^T \in \mathbb{R}^{1 \times m}$

$\frac{\partial J}{\partial b^{[2]}} = \frac{\partial J}{\partial o} \frac{\partial o}{\partial b^{[2]}} = \delta^{[1]}$

Chain rule for matrix multiplication - 1

$z = Wu + b$ (u: input var.)

$J = J(z)$

$u \in \mathbb{R}^d$

$W \in \mathbb{R}^{m \times d}, \quad b \in \mathbb{R}^d$

$\frac{\partial J}{\partial u} = W^T \cdot \frac{\partial J}{\partial z} \qquad \frac{\partial J}{\partial z} \in \mathbb{R}^m$

$$\frac{\partial J}{\partial W} = \frac{\partial J}{\partial W} \cdot u^T \qquad W^T \in \mathbb{R}^{d \times m}, \quad \frac{\partial J}{\partial u} \in \mathbb{R}^d$$

two layer NN case

$$\frac{\partial J}{\partial W^{[2]}} = \frac{\partial J}{\partial z} \cdot x^T$$

$W^{[2]} \leftrightarrow W$
$J \leftrightarrow J$
$z \leftrightarrow z$
$x \leftrightarrow u$
$b^{[2]} \leftrightarrow b$

$$\frac{\partial J}{\partial b^{[2]}} = \frac{\partial J}{\partial z} \cdot \frac{\partial z}{\partial b^{[2]}} = \frac{\partial J}{\partial z}$$

Chain rule for matrix multiplication - 2

$a = \sigma(z) \qquad a, z \in \mathbb{R}^m$

$J = J(a) \qquad \sigma$ is element-wise application

$$\frac{\partial J}{\partial z} = \frac{\partial J}{\partial a} \odot \sigma'(z)$$

$$\frac{\partial J}{\partial z} = \frac{\partial J}{\partial a} \odot \sigma'(z)$$

$$= \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial a} \odot \sigma'(z)$$

$$= \frac{\partial J}{\partial o} \cdot (W^{[2]})^T \odot \sigma'(z) \qquad \left(\frac{\partial J}{\partial}^{[1]} \text{ from above}\right)$$

Multiple layer example

$$a^{[1]} = \text{relu}(W^{[1]} x + b^{[1]})$$

$$a^{[2]} = \text{relu}(W^{[2]} a^{[1]} + b^{[2]})$$

$$\vdots$$

$$a^{[r]} \Rightarrow h_\theta(x) = W^{[r]} a^{[r-1]} + b^{[r]}$$

$$J = \frac{1}{2}(a^{[r]} - y)^2$$

#Note $\delta^{[k]} = \frac{\partial J}{\partial z^{[k]}} \in \mathbb{R}^{m_k}$ if $a^{[k]} \in \mathbb{R}^{m_k}$

$$\delta^{[r]} = \frac{\partial J}{\partial z^{[r]}} = (a^{[r]} - y) = (z^{[r]} - y)$$

$$\delta^{[r-1]} = \frac{\partial J}{\partial z^{[r-1]}} = \left(w^{[r]^T} \delta^{[r]}\right) \odot \text{relu}'(z^{[r-1]})$$

$$\frac{\partial J}{\partial z^{[r-1]}} = \frac{\partial J}{\partial z^{[r]}} \cdot \frac{\partial z^{[r]}}{\partial a^{[r-1]}} \cdot \frac{\partial a^{[r-1]}}{\partial z^{[r-1]}}$$

$$\vdots$$

$$\delta^{[k]} = \frac{\partial J}{\partial z^{[k]}} = \left(w^{[k+1]^T} \delta^{[k+1]}\right) \odot \text{relu}'(z^{[k]})$$

$$\Rightarrow \frac{\partial J}{\partial w^{[k+1]}} = \delta^{[k+1]} \cdot z^{[k]^T}$$

$$\boxed{z^{[0]} = x}$$

$$\vdots$$

$$\frac{\partial J}{\partial b^{[k+1]}} = \delta^{[k+1]}$$