1. MLP w/ one hidden layer. x and y are one-hot vectors representing words.

Let $h = W_{embed} x$, $s = W_{unembed} h$, $p = softmax(s)$, $p(y_c = 1 | s) = p_c$

a) Let $size(x) = size(y) = N$ and Let $size(h) = M$

$$\boxed{\begin{array}{l} W_{embed} \in \mathbb{R}^{M \times N} \\ W_{unembed} \in \mathbb{R}^{N \times M} \end{array}}$$

The columns of W_{embed} each represent the specific embedding information for each word represented by the one-hot vector.

b)

$$\begin{aligned} J &= \log p(y|s) \\ &= \sum_c y_c \log p_c \\ &= \sum_c y_c (s_c - \log z) \\ &= \sum_c y_c s_c - \log z \cancelto{1}{\sum_c y_c} \\ &= \left(\sum_c y_c s_c\right) - \log z \end{aligned}$$

$p(y|s) = \prod_c p_c^{y_c}$

$p_c = \frac{e^{s_c}}{Z}$ where $Z = \sum_c e^{s_c}$

$$\begin{aligned} \frac{\partial J}{\partial s_k} &= y_k - \frac{1}{Z} \frac{\partial Z}{\partial s_k} \\ &= y_k - \frac{1}{Z} e^{s_k} \\ &= y_k - p_k \end{aligned}$$

$$\boxed{\frac{\partial J}{\partial s} = y - p = \text{error}}$$

$$\begin{aligned} \frac{\partial J}{\partial h_j} &= \frac{\partial J}{\partial s_k} \frac{\partial s_k}{\partial h_j} \\ &= \frac{\partial J}{\partial s_k} W^u_{kj} \end{aligned}$$

$$\boxed{\frac{\partial J}{\partial h} = W_{unembed}^T \frac{\partial J}{\partial s}}$$

$S_k = \sum_j W^u_{kj} h_j$

let $W^u = W_{unembed}$

$$\begin{aligned} \frac{\partial J}{\partial W^u_{kj}} &= \frac{\partial J}{\partial s_k} \frac{\partial s_k}{\partial W^u_{kj}} \\ &= \frac{\partial J}{\partial s_k} \cdot h_j \end{aligned}$$

$$\boxed{\frac{\partial J}{\partial W^u} = \frac{\partial J}{\partial s} h^T}$$

$$\frac{\partial J}{\partial W^e_{jk}} = \frac{\partial J}{\partial h_j}\frac{\partial h_j}{\partial W^e_{jk}}$$

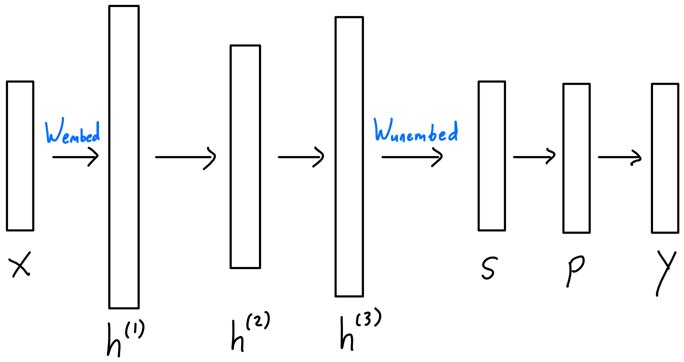
$$= \frac{\partial J}{\partial h_j} x_k$$

$$\boxed{\frac{\partial J}{\partial W^e} = \frac{\partial J}{\partial h} x^T}$$

$h_j = \sum_k W^e_{jk} x_k$

let $W^e = W_{embed}$

c)



2. let x_t be one-hot vectors. $x_1 =$ "I" $x_2 =$ "love" $x_3 =$ "machine"

let $h_0 = 0$, $h_t = \tanh(W^e x_t + W^r h_{t-1})$ for $t = 1, 2, 3$

let $s = W^u h_3$, $p = \text{softmax}(s)$, $x_4 \sim p$

a)

$$\begin{tikzcd}
 & & & s \arrow[r] & p \arrow[r] & x_4 \\
h_0 \arrow[r] & h_1 \arrow[r, "W^r"] & h_2 \arrow[r] & h_3 \arrow[u, "W^u"'] & & \\
 & x_1 \arrow[u, "W^e"'] & x_2 \arrow[u] & x_3 \arrow[u] & &
\end{tikzcd}$$

let $W^r = W_{recurrent}$

$W^e = W_{embed}$

$W^u = W_{unembed}$

b) calculating $\frac{\partial J}{\partial s}, \frac{\partial J}{\partial W^u}, \frac{\partial J}{\partial h_3}$:

$$\frac{\partial J}{\partial s} = y - p = \text{error}$$

$$\boxed{\frac{\partial J}{\partial W^u} = \frac{\partial J}{\partial s} h_3^T}$$

$$\frac{\partial J}{\partial h_3} = W_{unembed}^T \frac{\partial J}{\partial s}$$

using results from problem 1

calculating $\frac{\partial J}{\partial h_2}, \frac{\partial J}{\partial h_1}, \frac{\partial J}{h_0}$:

$$\frac{\partial J}{\partial h_{t-1,i}} = \frac{\partial J}{\partial h_{t,j}} \frac{\partial h_{t,j}}{\partial g_{t,j}} \frac{\partial g_{t,j}}{\partial h_{t-1,i}}$$

$$h_{tj} = \tanh\Big(\overbrace{\sum_k W^{e,T}_{jk} x_{t,k} + \sum_i W^{r,T}_{ji} h_{t-1,i}}^{g_{t,j}}\Big)$$

$$= \frac{\partial J}{\partial h_{t_j}} (1-h_{t_j}^2) W^r_{ji}$$

$$\frac{\partial J}{\partial h_{t-1}} = W^{r,T} \, \text{diag}(\mathbb{1} - h_t \odot h_t) \frac{\partial J}{\partial h_t}$$

property: $\frac{\partial}{\partial z} \tanh(z) = \text{diag}(\mathbb{1} - \tanh(z)^2)$

$$\frac{\partial h_t}{\partial g_t} = \text{diag}(\mathbb{1} - h_t \odot h_t)$$

calculating $\frac{\partial J}{\partial W^r}$

$$\begin{aligned}
\frac{dJ}{dW^r} &= \frac{\partial J}{\partial s} \frac{ds}{dW^r} \\
&= \frac{\partial J}{\partial s} \frac{\partial s}{\partial h_3} \frac{dh_3}{dW^r} \\
&= \frac{\partial J}{\partial h_3} \frac{\partial h_3}{\partial g_3} \frac{dg_3}{dW^r} \\
&= \frac{\partial J}{\partial h_3} \frac{\partial h_3}{\partial g_3} \left(\frac{\partial g_3}{\partial W^r} + \frac{\partial g_3}{\partial h_2} \frac{dh_2}{dW^r} \right) \\
&= \frac{\partial J}{\partial h_3} \frac{\partial h_3}{\partial g_3} \frac{\partial g_3}{\partial W^r} + \frac{\partial J}{\partial h_2} \frac{dh_2}{\partial W^r} \\
&= \quad \ldots \\
&= \frac{\partial J}{\partial h_3} \frac{\partial h_3}{\partial g_3} \frac{\partial g_3}{\partial W^r} + \frac{\partial J}{\partial h_2} \frac{\partial h_2}{\partial g_2} \frac{\partial g_2}{\partial W^r} + \frac{\partial J}{\partial h_1} \frac{\partial h_1}{\partial g_1} \frac{dg_1}{dW^r} \\
&= \boxed{\text{diag}(\mathbb{1} - h_3 \odot h_3) \frac{\partial J}{\partial h_3} h_2^T + \text{diag}(\mathbb{1} - h_2 \odot h_2) \frac{\partial J}{\partial h_2} h_1^T + \text{diag}(\mathbb{1} - h_1 \odot h_1) \frac{\partial J}{\partial h_1} h_0^T}
\end{aligned}$$

$h_t = g(W, h_{t-1}(W))$

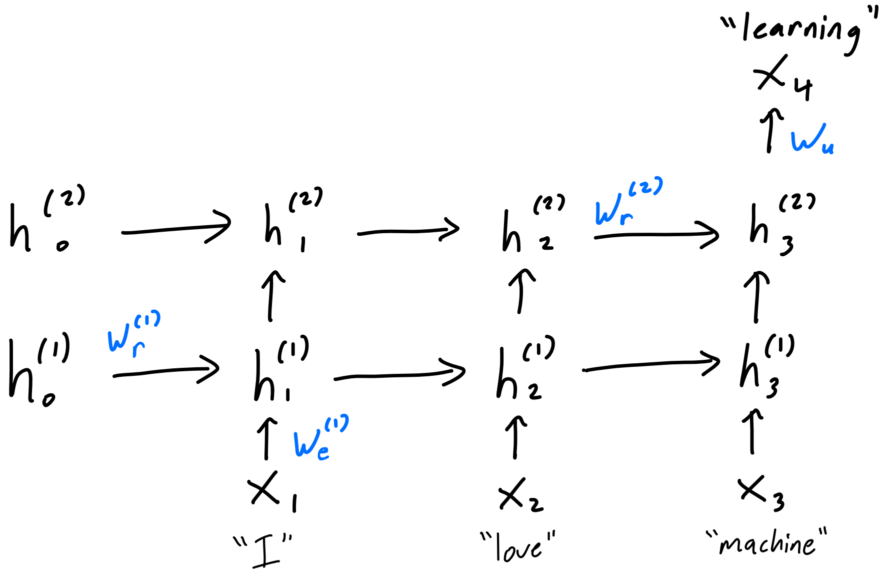
$g_t = W^e x_t + W^r h_{t-1}$

calculating $\frac{\partial J}{\partial W^e}$

using similar logic to $\frac{\partial J}{\partial W^r}$,

$$\begin{aligned}
\frac{dJ}{dW^e} &= \frac{\partial J}{\partial h_3} \frac{\partial h_3}{\partial g_3} \frac{\partial g_3}{\partial W^e} + \frac{\partial J}{\partial h_2} \frac{\partial h_2}{\partial g_2} \frac{\partial g_2}{\partial W^e} + \frac{\partial J}{\partial h_1} \frac{\partial h_1}{\partial g_1} \frac{dg_1}{dW^e} \\
&= \boxed{\text{diag}(\mathbb{1} - h_3 \odot h_3) \frac{\partial J}{\partial h_3} x_3^T + \text{diag}(\mathbb{1} - h_2 \odot h_2) \frac{\partial J}{\partial h_2} x_2^T + \text{diag}(\mathbb{1} - h_1 \odot h_1) \frac{\partial J}{\partial h_1} x_1^T}
\end{aligned}$$

3.



1. $h_t = \tanh(\overbrace{W_r h_{t-1} + W_e x_t}^{g_t})$ let $h_t \in \mathbb{R}^n$, $h_{t-1} \in \mathbb{R}^m$, $W \in \mathbb{R}^{n \times m}$

$$\frac{\partial J}{\partial h_{t-1}} = \frac{\partial J}{\partial h_t} \frac{\partial h_t}{\partial g_t} \frac{d g_t}{d h_{t-1}}$$

$$= \underbrace{\frac{\partial J}{\partial h_t}}_{n \times 1} \cdot \underbrace{\mathrm{diag}(\mathbb{1} - \tanh^2 g_t)}_{n \times n} \cdot \underbrace{W_r}_{n \times m}$$

reorder so multiplication makes sense

$$= W_r^T \, \mathrm{diag}(\mathbb{1} - \tanh^2 g_t) \frac{\partial J}{\partial h_t}$$

$$\frac{\partial J}{\partial h_1} = W_r^T \, \mathrm{diag}(\mathbb{1} - \tanh^2 g_2) \frac{\partial J}{\partial h_2}$$

$$= \boxed{W_r^T \, \mathrm{diag}(\mathbb{1} - \tanh^2 g_2) \, W_r^T \, \mathrm{diag}(\mathbb{1} - \tanh^2 g_3) \frac{\partial J}{\partial h_3}}$$

2. $h_t = \tanh(\overbrace{W_r h_{t-1} + W_e x_t}^{g_t}) + h_{t-1}$

$$\frac{\partial J}{\partial h_{t-1}} = \frac{\partial J}{\partial h_t} \frac{\partial h_t}{\partial h_{t-1}}$$

$$= \frac{\partial J}{\partial h_t} \left[\frac{\partial \tanh(g_t)}{\partial g_t} \frac{d g_t}{d h_{t-1}} + I \right]$$

$$= \frac{\partial J}{\partial h_t} \left[\mathrm{diag}(\mathbb{1} - \tanh^2 g_t) W_r + I \right]$$

reorder

$$= \left[W_r^T \, \mathrm{diag}(\mathbb{1} - \tanh^2 g_t) + I \right] \frac{\partial J}{\partial h_t}$$

$$\frac{\partial J}{\partial h_1} = \left[W_r^T \, \mathrm{diag}(\mathbb{1} - \tanh^2 g_2) + I \right] \frac{\partial J}{\partial h_2}$$

$$= \boxed{\left[W_r^T \, \mathrm{diag}(\mathbb{1} - \tanh^2 g_2) + I \right] \left[W_r^T \, \mathrm{diag}(\mathbb{1} - \tanh^2 g_3) + I \right] \frac{\partial J}{\partial h_3}}$$

(2) alleviates the vanishing gradient problem because even when $\mathbb{1} - \tanh^2 g_t$ goes to 0 (when g_t is large), $\frac{\partial J}{\partial h_{t-1}}$ will still have the $\frac{\partial J}{\partial h_t}$ term from the skip connection to keep it nonzero

In both cases, $\frac{\partial J}{\partial h_{t-1}}$ is a function of $\frac{\partial J}{\partial h_t}$.
Thus, $\frac{\partial J}{\partial h_3}$ is needed for $\frac{\partial J}{\partial h_2}$, which is needed for $\frac{\partial J}{\partial h_1}$