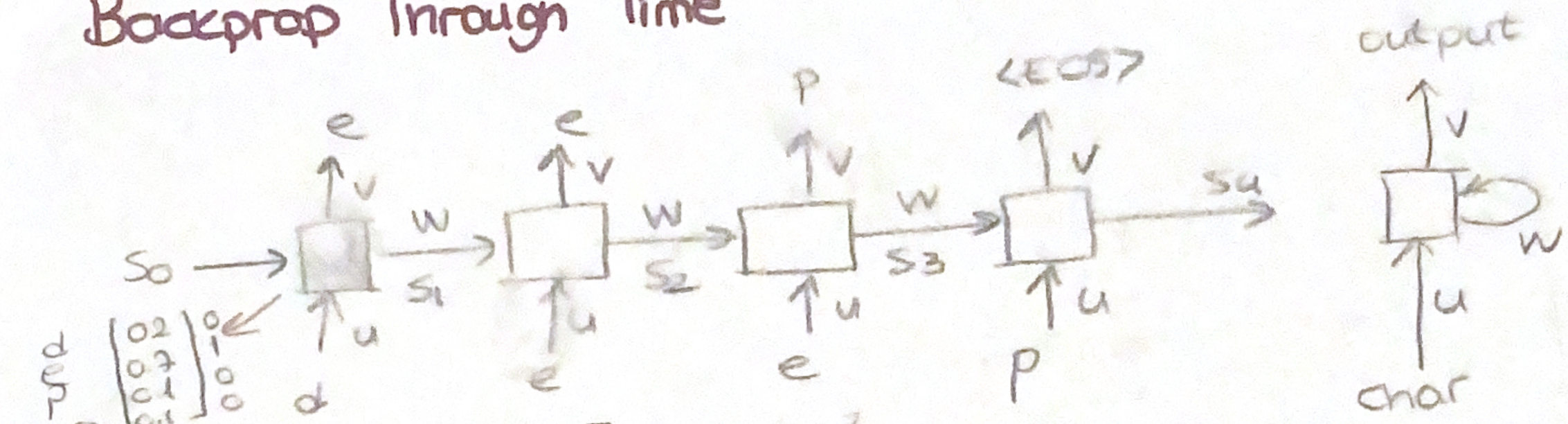
Backprop Through Time



$$\text{Total Loss} \rightarrow \sum_{t=1}^{T} \mathcal{L}_t(\theta)$$ } sum of loss over all steps

$\mathcal{L}_t(\theta) = -\log(y_{tc})$
↳ prob produced for true class at time step t (0.7 in above example, log will decrease if it gets closer to 1)

For backprop we compute gradients for W, U, V

$$\text{For } V: \quad \frac{\partial \mathcal{L}(\theta)}{\partial V} = \frac{\partial \mathcal{L}_1(\theta)}{\partial V} + \frac{\partial \mathcal{L}_2(\theta)}{\partial V} + \cdots$$ } standard backprop

this is affected by states

V is a matrix

s_4 depends on s_3 & w
s_3 depends on s_2 & w

$$\text{For } W: \quad \frac{\partial \mathcal{L}(\theta)}{\partial W}$$

$$\frac{\partial \mathcal{L}_4(\theta)}{\partial W} = \frac{\partial \mathcal{L}_4(\theta)}{\partial s_4} \cdot \frac{\partial s_4}{\partial W}$$

this is calculated with regular BP

we can't calculate this
$s_4 = \sigma(Ws_3 + b)$

Total derivative has 2 parts.

Explicit → $\frac{\partial^+ s_4}{\partial W}$ } treats all inputs as constant

Implicit → Sum all indirect paths from s_4 to W.

$$\frac{\partial s_4}{\partial W} = \frac{\partial^+ s_4}{\partial W} + \frac{\partial s_4}{\partial s_3} \cdot \frac{\partial s_3}{\partial W} = \frac{\partial^+ s_4}{\partial W} + \frac{\partial s_4}{\partial s_3}\left[\underbrace{\frac{\partial^+ s_3}{\partial W}}_{\text{exp}} + \frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial W}\right]$$

$$\cdots \quad \frac{\partial^+ s_2}{\partial s_1} \cdot \frac{\partial s_1}{\partial W}$$