Training RNNs

In RNN: Back Propagation Through Time (BPTT).

$$\bar{S}_t = \Phi(\bar{x}_t \cdot W_x + \bar{S}_{t-1} \cdot W_s)$$

e.g., $\bar{S}_t = \tanh(\bar{x}_t \cdot W_x + \bar{S}_{t-1} \cdot W_s)$

$$\bar{y}_t = \sigma(\bar{S}_t \cdot W_y)$$

σ → Softmax

$E_t = (\bar{d}_t - \bar{y}_t)^2$ error function is MSE

How does BPTT work?

$t = 3$

$$E_3 = (\bar{d}_3 - \bar{y}_3)^2$$

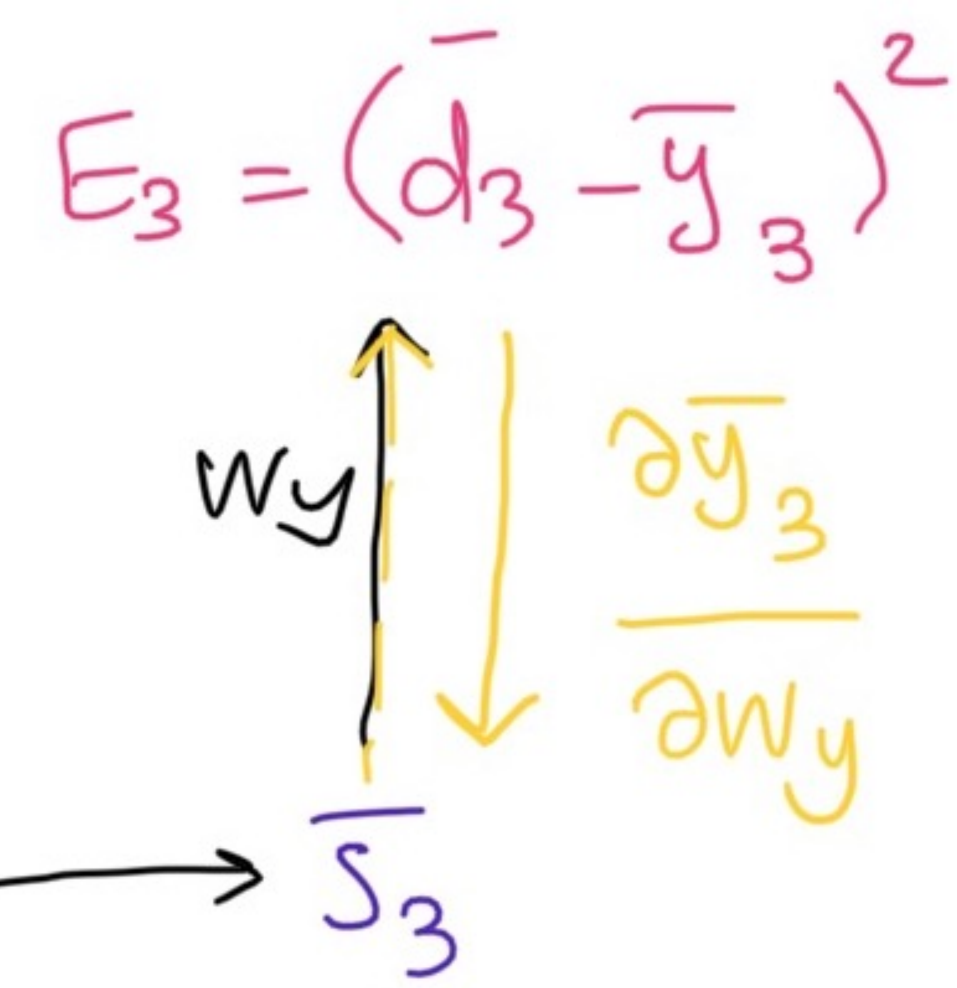
$\bar{d}_3$ → desired output

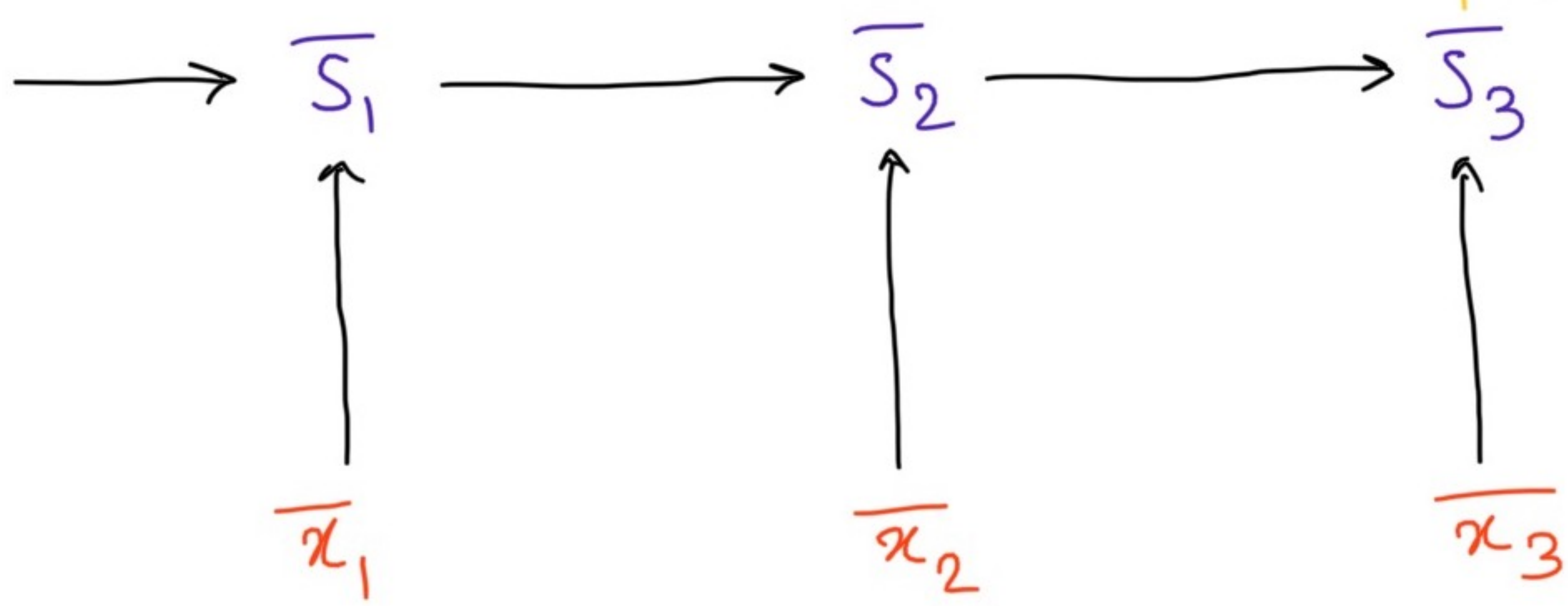
$\bar{y}_3$ → Calculated output

but we also need info from $t=2$ and $t=1$.

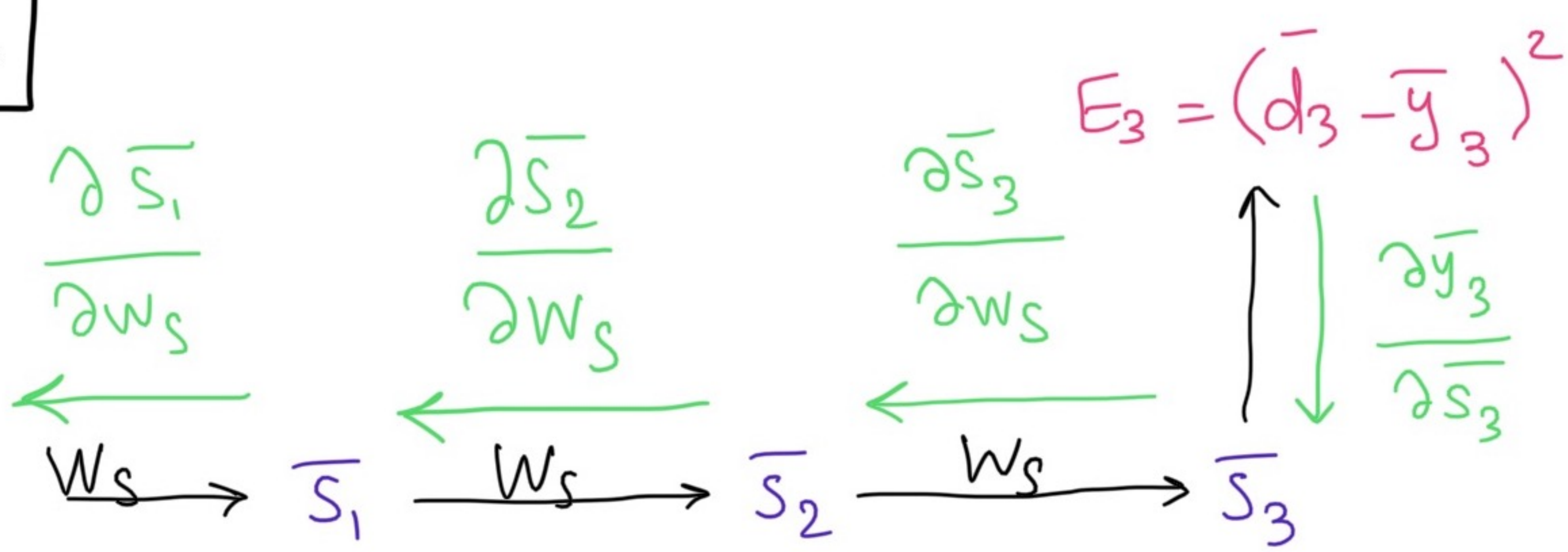
$\boxed{W_y}$

$$\boxed{\frac{\partial E_3}{\partial w_y} = \frac{\partial E_3}{\partial \bar{y}_3} \cdot \frac{\partial \bar{y}_3}{\partial w_y}}$$





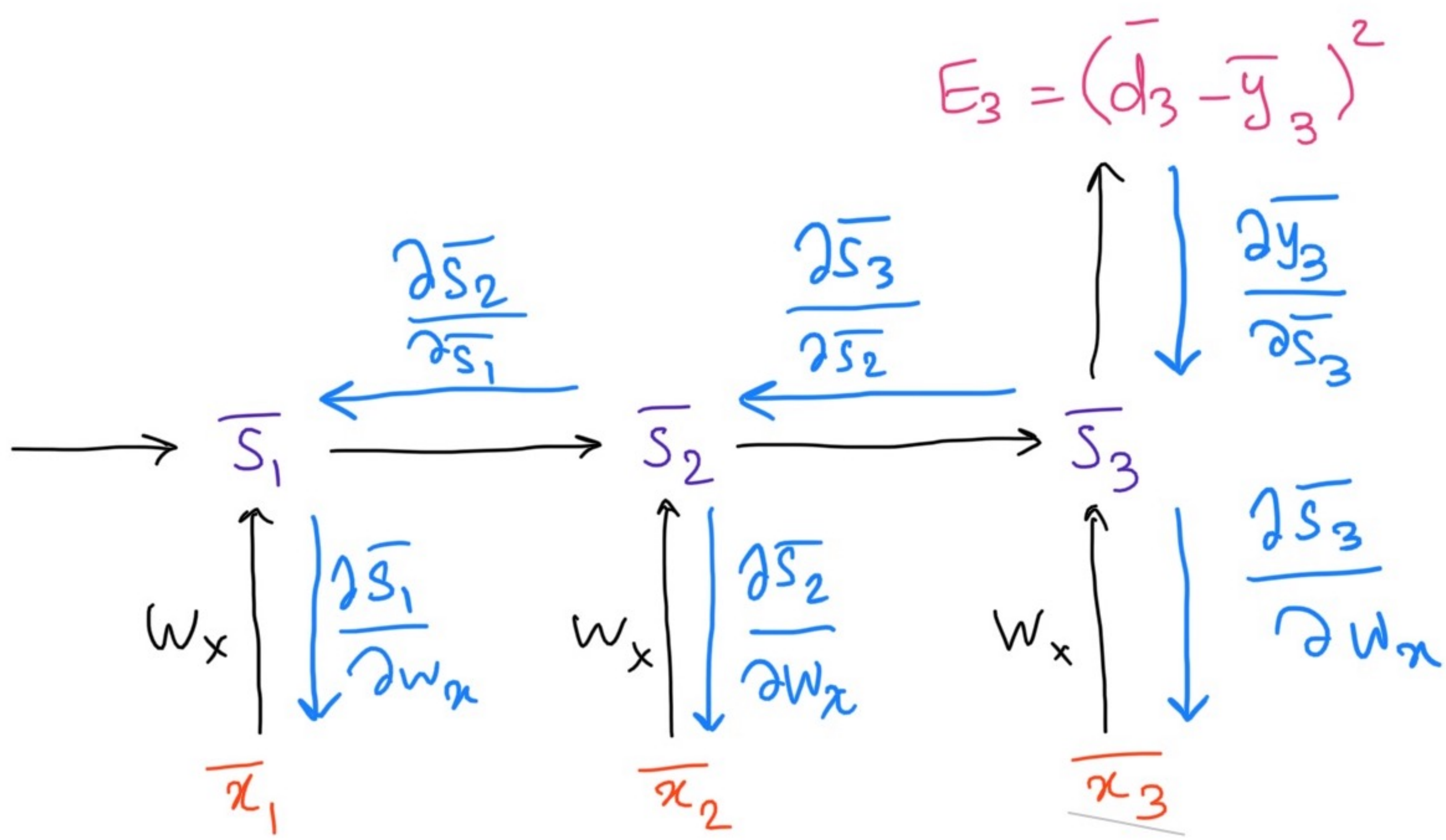
$\boxed{W_s}$



$$\begin{aligned}
\frac{\partial E_3}{\partial W_s} &= \frac{\partial E_3}{\partial \bar{y}_3} \cdot \frac{\partial \bar{y}_3}{\partial \bar{s}_3} \cdot \frac{\partial \bar{s}_3}{\partial w_s} \\
&+ \frac{\partial E_3}{\partial \bar{y}_3} \cdot \frac{\partial \bar{y}_3}{\partial \bar{s}_3} \cdot \frac{\partial \bar{s}_3}{\partial \bar{s}_2} \cdot \frac{\partial \bar{s}_2}{\partial w_s} \\
&+ \frac{\partial E_3}{\partial \bar{y}_3} \cdot \frac{\partial \bar{y}_3}{\partial \bar{s}_3} \cdot \frac{\partial \bar{s}_3}{\partial \bar{s}_2} \cdot \frac{\partial \bar{s}_2}{\partial \bar{s}_1} \cdot \frac{\partial \bar{s}_1}{\partial w_s}
\end{aligned}$$

More generally: $\dfrac{\partial E_N}{\partial W_S} = \sum_{i=1}^{N} \dfrac{\partial E_N}{\partial \bar{y}_N} \cdot \dfrac{\partial \bar{y}_N}{\partial \bar{s}_i} \cdot \dfrac{\partial \bar{s}_i}{\partial W_S}$

W_x



$$\frac{\partial E_3}{\partial W_x} = \frac{\partial E_3}{\partial \bar{y}_3} \cdot \frac{\partial \bar{y}_3}{\partial \bar{s}_3} \cdot \frac{\partial \bar{s}_3}{\partial w_x}$$

$$+ \frac{\partial E_3}{\partial \bar{y}_3} \cdot \frac{\partial \bar{y}_3}{\partial \bar{s}_3} \cdot \frac{\partial \bar{s}_3}{\partial \bar{s}_2} \cdot \frac{\partial \bar{s}_2}{\partial w_x}$$

$$+ \frac{\partial E_3}{\partial \bar{y}_3} \cdot \frac{\partial \bar{y}_3}{\partial \bar{s}_3} \cdot \frac{\partial \bar{s}_3}{\partial \bar{s}_2} \cdot \frac{\partial \bar{s}_2}{\partial \bar{s}_1} \cdot \frac{\partial \bar{s}_1}{\partial w_x}$$

More generally: $\dfrac{\partial E_N}{\partial W_x} = \sum_{i=1}^{N} \dfrac{\partial E_N}{\partial \bar{y}_N} \cdot \dfrac{\partial \bar{y}_N}{\partial \bar{s}_i} \cdot \dfrac{\partial \bar{s}_i}{\partial W_x}$

- for training: updating weights $\underbrace{\text{every N steps}}_{\text{mini-batch}}$

- what happens if we have too many steps?
  up to 8-10 steps, this BPTT works, but beyond that, the vanishing gradient problem happens. $\longrightarrow$ LSTM is born!

- Gradient clipping: → avoiding Exploding Gradient Problem.

At eatch timestep t:

$$\delta = \frac{\partial y}{\partial W_{ij}} > \text{threshold ?}$$

If so, normalize the gradient.