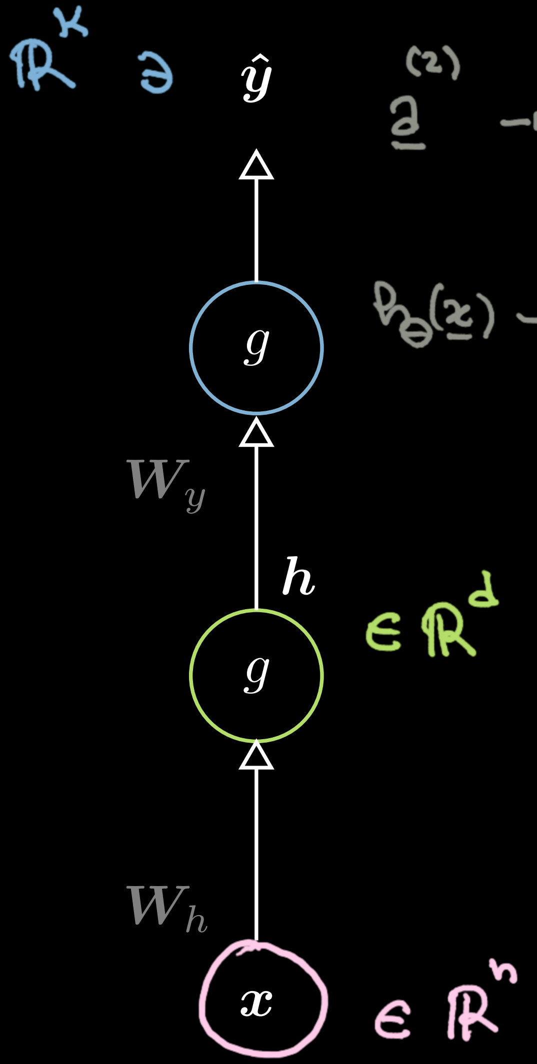
Practical 4.1

Recurrent Neural Networks – Forward and backward

Overview (I)

- Vanilla Neural Networks (NN) equations review
- Recurrent Neural Networks (RNN) equations
- Back-propagation through time (BPTT)

NN review



$$\underline{a}^{(2)} \to \underline{h} = g\Big(\overbrace{W_h\, \underline{x} + \underline{b}_h}^{\Theta^{(1)} \quad \hat{\underline{x}}}\Big) \qquad \sigma \downarrow$$

$$h_\Theta(\underline{x}) \to \hat{\underline{y}} = g\Big(\underbrace{W_y\, \underline{h} + \underline{b}_y}_{\Theta^{(2)} \quad \hat{\underline{a}}^{(2)}}\Big) \qquad \sigma \uparrow$$

$W_h \in \mathbb{R}^{d \times n}$

$\underline{b}_h \in \mathbb{R}^d$

$W_y \in \mathbb{R}^{K \times d}$

$\underline{b}_y \in \mathbb{R}^K$

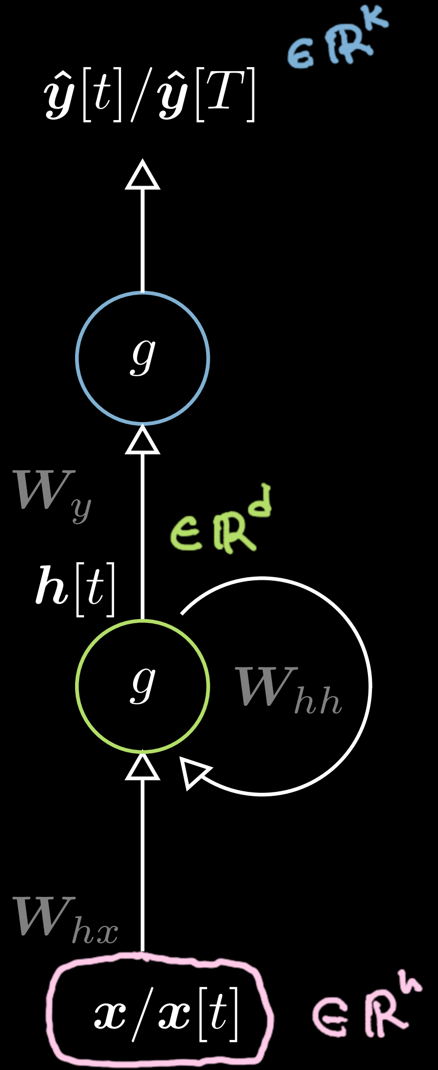
$g(\cdot) : \sigma(\cdot), \tanh(\cdot), \dots$

$\hat{\underline{y}} = \hat{\underline{y}}(\underline{x}) \qquad \hat{\underline{y}} : \mathbb{R}^n \to \mathbb{R}^K \qquad \underline{x} \mapsto \hat{\underline{y}}$

$\hat{\underline{y}} : \mathbb{R}^n \to \mathbb{R}^d \to \mathbb{R}^K, \quad d \gg n, K$

$$W = \begin{bmatrix} - \underline{w}^{(1)} - \\ - \underline{w}^{(2)} - \\ \vdots \end{bmatrix}$$

RNN Recurrent Neural Network



$W_{hx} \in \mathbb{R}^{d \times n}$ $\quad W_{hh} \in \mathbb{R}^{d \times d}$

$W_y \in \mathbb{R}^{k \times d}$

$\underline{b}_h \in \mathbb{R}^d$ $\quad \underline{b}_y \in \mathbb{R}^k$

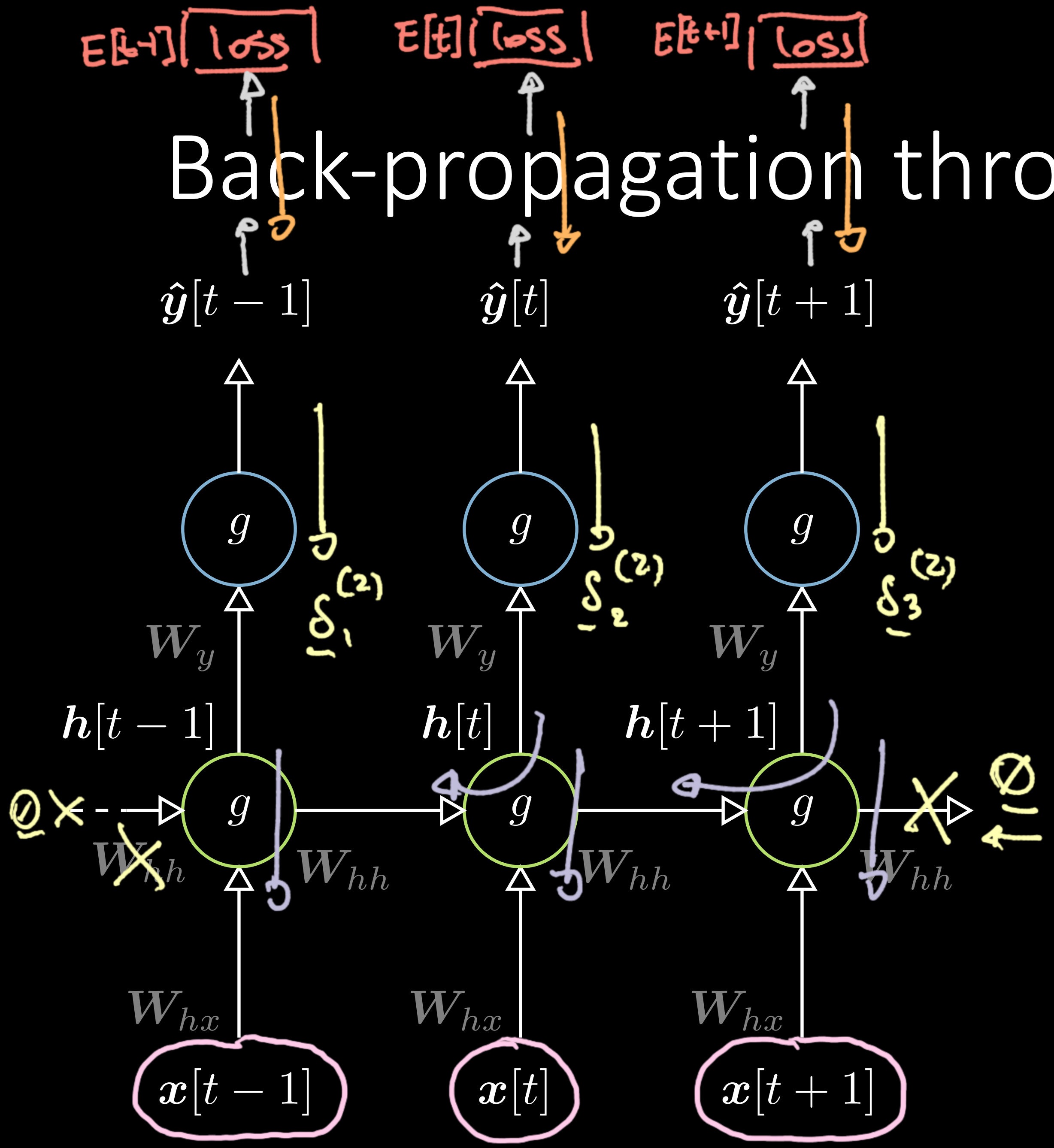
$$\underline{h}[t] = g(W_{hx}\, \underline{x}[t] + W_{hh}\, \underline{h}[t-1] + \underline{b}_h)$$

$$\underline{h}[0] := \underline{0}$$

$$\underline{\hat{y}}[t] = g(W_y\, \underline{h}[t] + \underline{b}_y)$$

$$\underline{h}[t] = g\left(W_h \begin{bmatrix} \underline{x}[t] \\ \underline{h}[t-1] \end{bmatrix} + \underline{b}_h\right) \qquad W_h = [W_{hx}\ W_{hh}] \in \mathbb{R}^{d \times (n+d)}$$

Back-propagation through time (BPTT)



$\{E_1, E_2, E_3\} \rightarrow \{\nabla_y E_1, \nabla_y E_2, \nabla_y E_3\}$

- For training we fix T train interval

$\{y[1], y[2], y[3]\}$

- ∅ init. state
- ∅ future knowledge

○ ○ ○ share same Θ, $\frac{\partial E}{\partial \Theta}$

$\Theta = \{\Theta^{(1)}, \Theta^{(2)}\}$ shared across temporal replica

○ ○ ○ share same parameters Θ and $\frac{\partial E}{\partial \Theta}$

$\{x[1], x[2], x[3]\}$