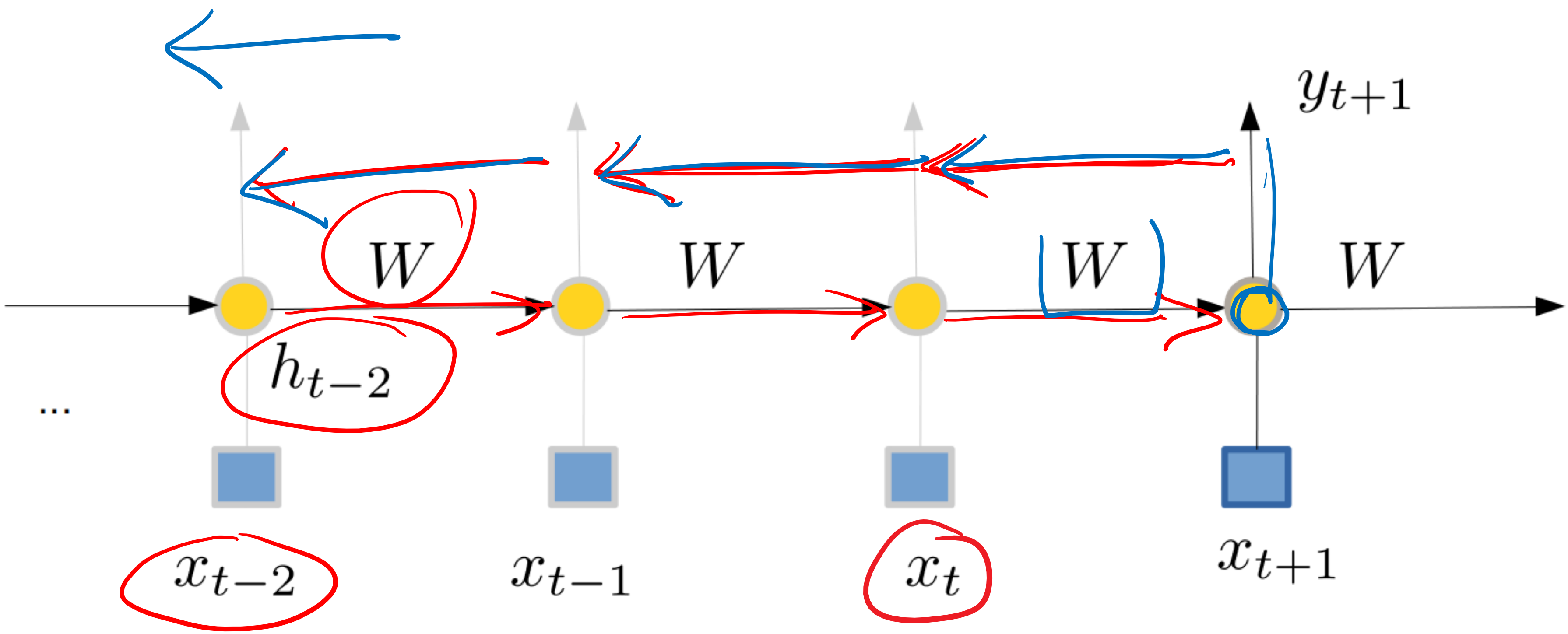
Understanding Vanishing Gradients

Heni Ben Amor, Ph.D.
Assistant Professor
Arizona State University

Ira A. Fulton Schools of Engineering
Arizona State University

Deeper Look at Vanishing Gradients

- We focus on the temporal connections



- **What happens when W is smaller than 1?**

Sigmoid Derivative

$\sigma'(z)$ $\sigma'(z)$

t-2 t-1

Sigmoid Activation

$$\sigma = \frac{1}{1 - e^{-x}}$$

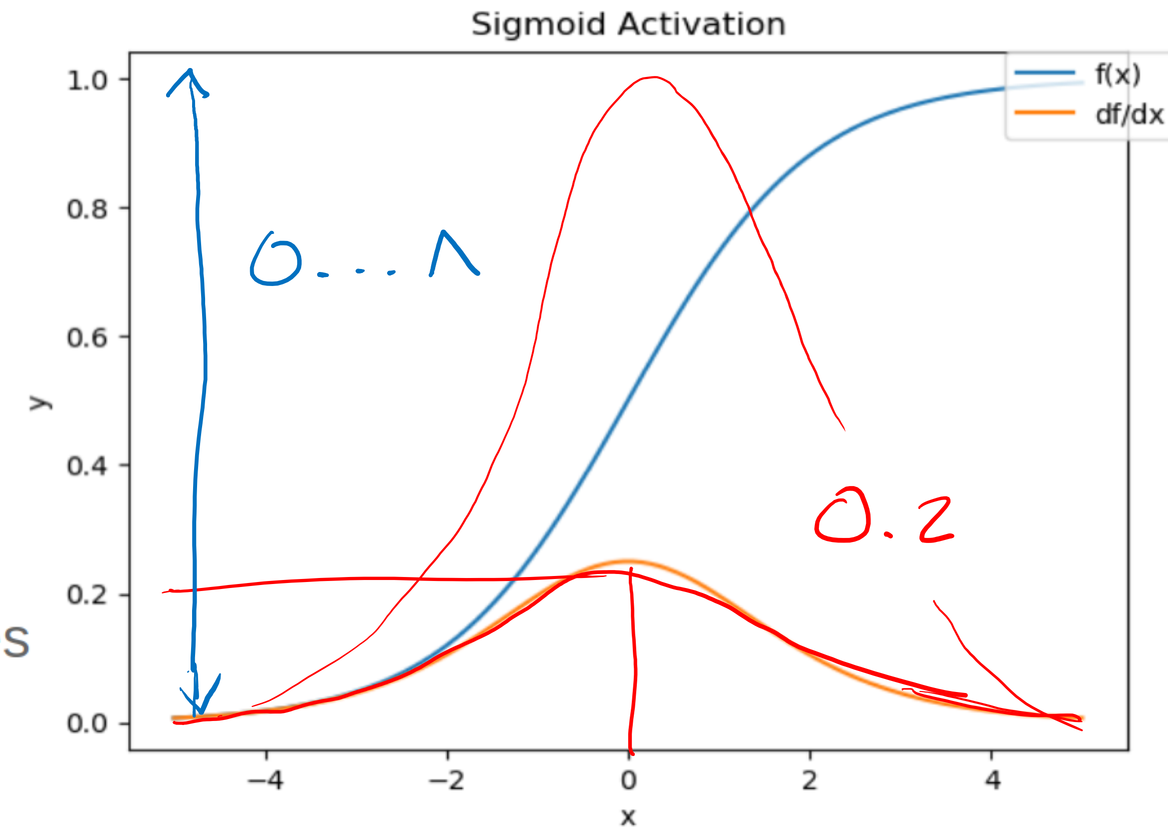
$$\sigma' = (1 - \sigma)\sigma$$

Max value of df/dx is .25

Temporal gradient vanishes quickly with this activation function

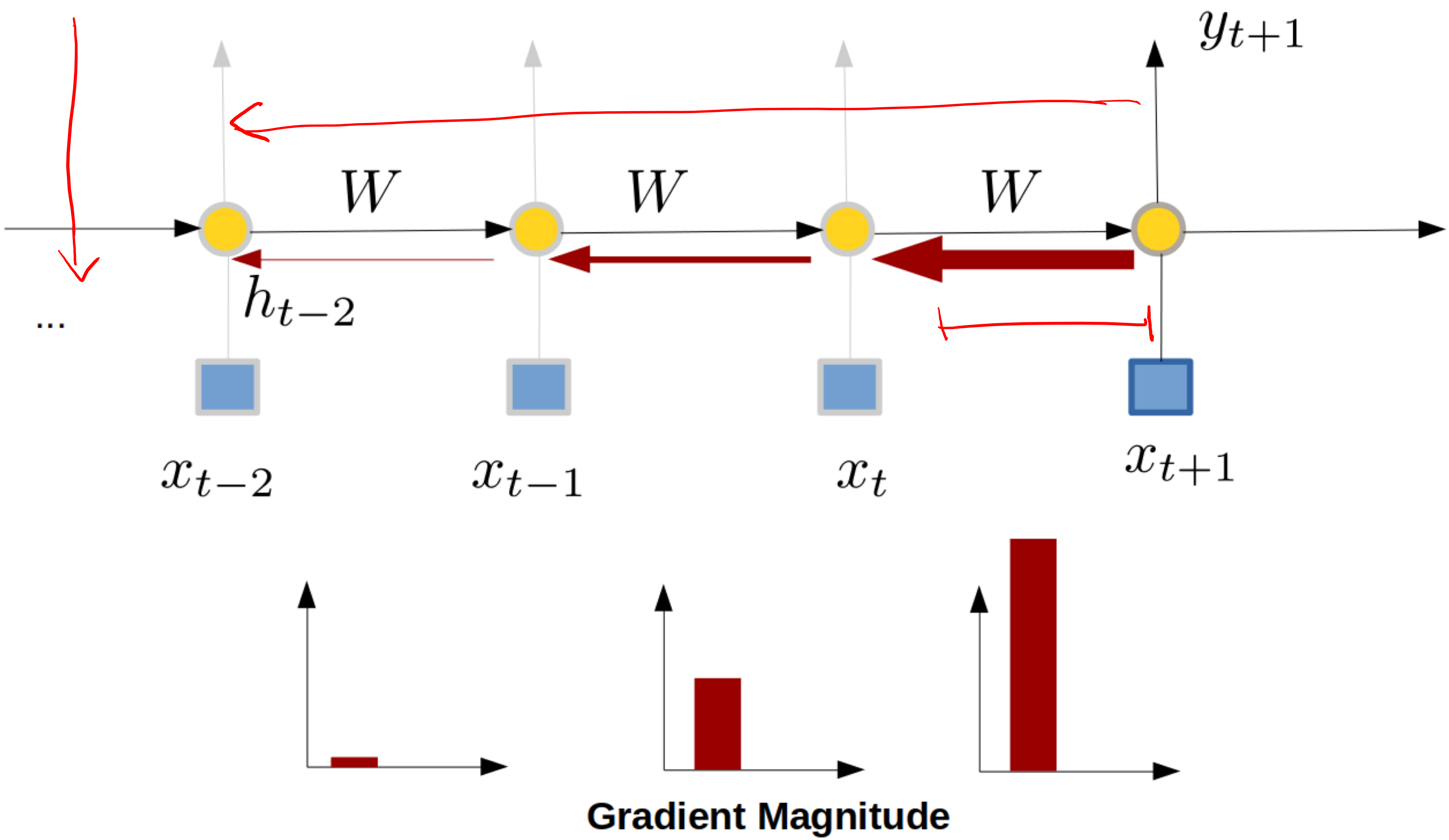
$$.25^2 = .0625$$

$$.25^5 = 0.00097$$



Vanishing Gradients Visualized

past timesteps have smaller influence



Theory Behind Vanishing Gradient

$$\nabla E = \left(\frac{\partial E}{\partial w_{11}}, \frac{\partial E}{\partial w_{12}}, \ldots \right)$$

- BPTT: calculate gradient and propagate through time (TT)

$$\theta = w_{ij} \quad \frac{\partial E}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}_t}{\partial \theta}$$

variables / weights of RNN

- Apply chain rule to incorporate hidden state

$$\frac{\partial \mathcal{L}_t}{\partial \theta} = \sum_{k=1}^{T} \left(\frac{\partial \mathcal{L}_t}{\partial h_t} \odot \boxed{\frac{\partial h_t}{\partial h_k}} \odot \frac{\partial h_k}{\partial \theta} \right)$$

- Temporal connections introduce dangerous product

$$\boxed{\frac{\partial h_t}{\partial h_k}} = \prod_{i=k}^{t} \frac{\partial h_i}{\partial h_{i-1}}$$

Vanishing and Exploding Gradients

- Challenge when training RNNs
- Gradients quickly shrink to negligible values
- Or, gradients may grow substantially and make learning unstable
- An immediate result of the **temporal connections**
- **Exponential growth in hidden state values**
- Effect: learning is slow and yields poor results