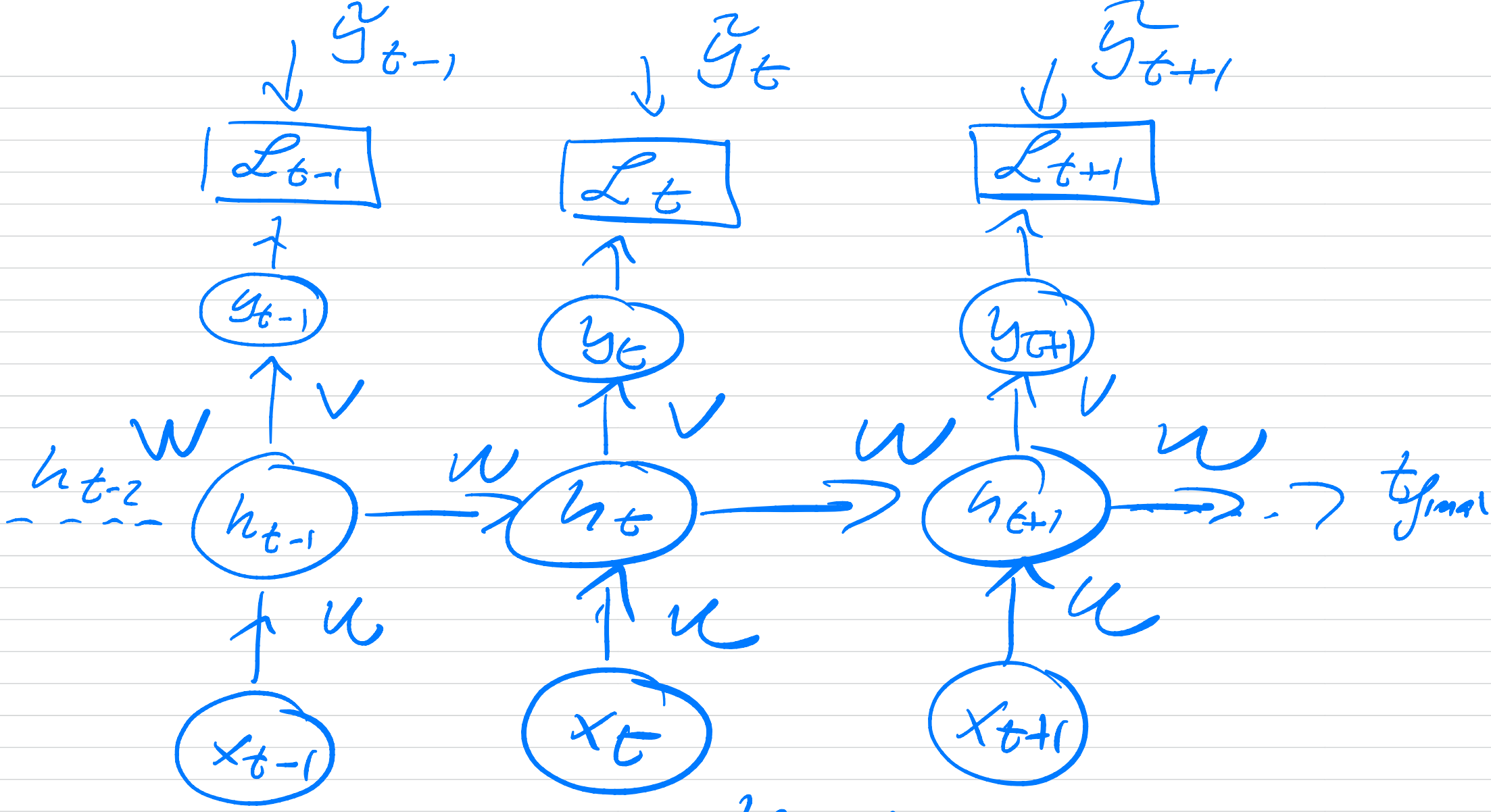
Lecture FYS5429, March 5, 2024

FYS 5429, MARCH 5, 2029

set of parameters

$$\theta = \{ u, W, V, b, c \}$$



$$\mathcal{L}(\epsilon) = \sum_{t=0}^{t_{final}} \mathcal{L}_t$$

$$z_t = U x_t + W h_{t-1} + b_t$$

$$h_t = \sigma_h(z_t)$$

$$r_t = V h_t + c_t \quad \leftarrow \text{bias}$$

$$y_t = \sigma_y(r_t)$$

$$\mathcal{L}_t(y_t, \tilde{y}_t)$$

Common strategies

- weight sharing
- truncation of number of steps where different weights are training

- Training is done iteratively
  - Feed forward pass (randomly assigned weights & Biases)
  - Back propagation in time (BPTT)

BPTT

$$\mathcal{L} = \sum_t \mathcal{L}_t$$

$\nabla_c \mathcal{L}$

$\nabla_b \mathcal{L}$, $\nabla_W \mathcal{L}$, $\nabla_V \mathcal{L}$

$\nabla_U \mathcal{L}$

can give rise to exploding or vanishing gradients.

$$\frac{\partial \mathcal{L}_t}{\partial W} = \frac{\partial \mathcal{L}_t}{\partial y_t} \frac{\partial y_t}{\partial h_t}$$

$$\times \left\{ \sum_{k=1}^{n} \left[\frac{\partial h_t}{\partial h_k} \right] \frac{\partial h_k}{\partial w} \right\}$$

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

can lead to exploding or vanishing gradients

assume same W for all steps

$$h_1 = W h_0$$

$$h_i = W^i h_0 = W h_{i-1}$$

$$= W \cdot W h_{i-2}$$

assume we can diagonalize W

$$W = U D U^T \qquad U U^T = U^T U = 1$$

eigenpairs are (λ_i, w_i)

$$h_0 = \sum_i \alpha_i w_i$$

$$W h_0 = \sum_i \alpha_i W w_i$$

$$= \sum_i \alpha_i \lambda_i w_i$$

$$W^t h_0 = \sum_i \lambda_i^t \alpha_i w_i$$

$$\lambda_0 > \lambda_1 > \lambda_2 > \ldots > \lambda_n$$

$\lambda_0 \geq 1$

$$h^t = W^t h_0 \approx \lambda_0 \alpha_0 w_0$$

$\lambda_0 > 1 \Rightarrow$ exploding gradients

$\lambda_0 = 1$, ideal case

$\lambda_0 < 1$ $\Rightarrow$ vanishing gradients

exploding gradient,
gradient clipping

gradient $\vec{g}$

if $\|\vec{g}\|_2 \geq \varepsilon$

$$\vec{g} \leftarrow \frac{\varepsilon}{\|\vec{g}\|_2} \vec{g}$$

endif