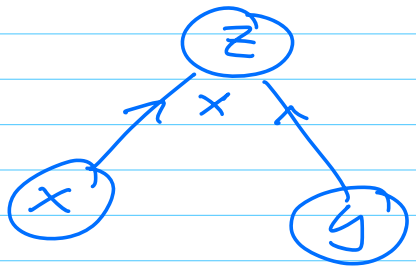
FYS 5429, MARCH 8, 2023

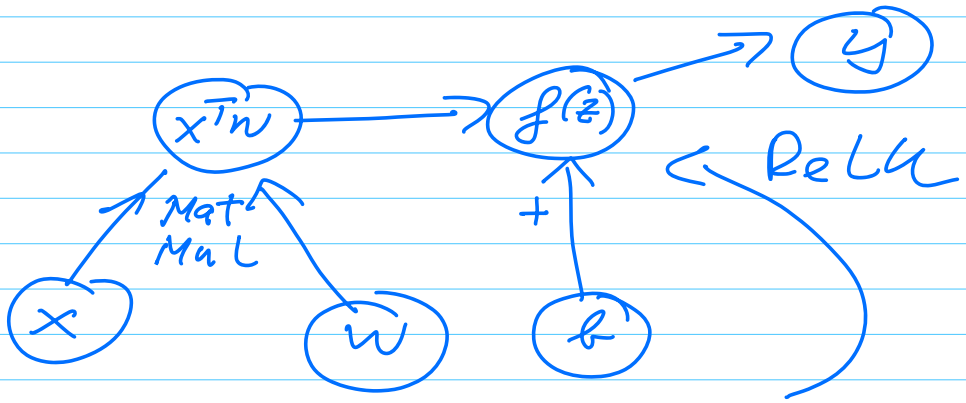
RNN

Graphical representations

$$z = x \cdot y$$



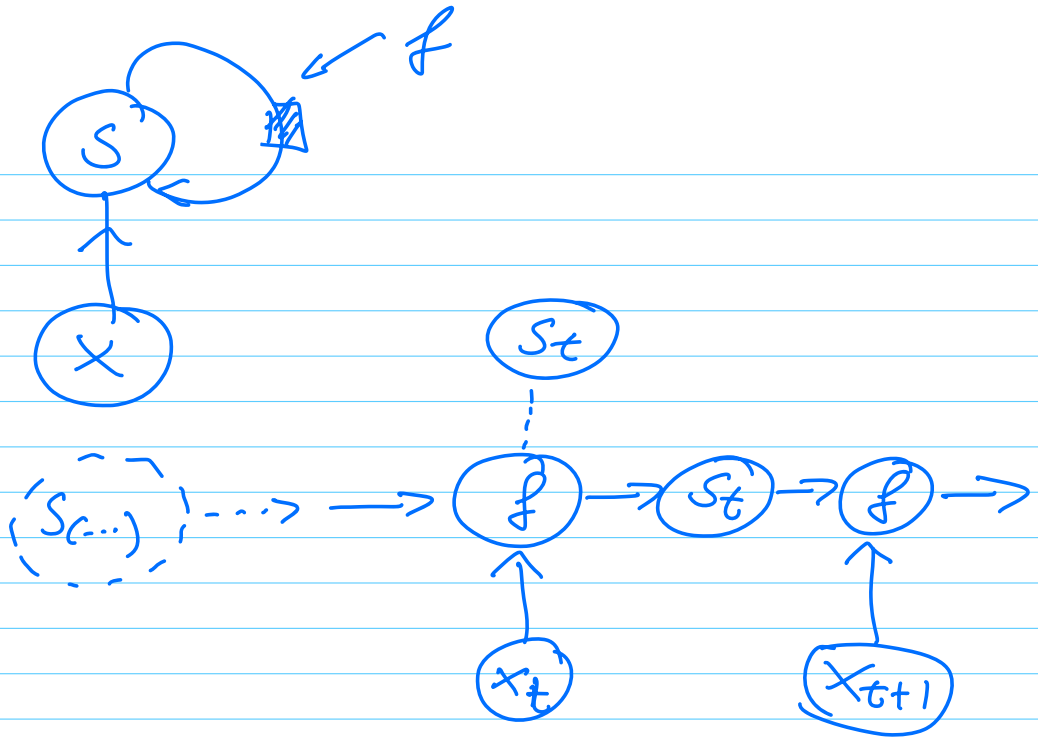
$$y = f(x^T W + b) = f(z)$$



$$y = \max(0, x^T W + b) = f(z)$$

Dynamical system, driven by an external signal x_t

$$s_t = f(s_{t-1}, x_t; \theta)$$



→ (StH) - - - → final stage.

RNNs contain at each stage
- new input (St-1, Xt)
- manipulate the state (f)
- reuses weights
- produces new output

Project 1 and RNNs

$$m\frac{d^2x}{dt^2} + \mu\frac{dx}{dt} + x(t) = F(t)$$

$$x_0 = x(t_0) \quad \wedge \quad v_0 = v(t_0)$$

$$v(t) = \frac{dx}{dt}$$

$$a = \frac{dv}{dt} = \frac{d^2x}{dt^2}$$

$$= -\frac{\mu}{m} v - \frac{x}{m} + F$$

$x \Rightarrow x_i' = x(t_i)$

$v \Rightarrow v_i = v(t_i)$

$x_{i+1} = x(t_i' + \Delta t)$

$t \Rightarrow t_i' = t_0 + i' \Delta t$

$i = 0, 1, 2, \ldots m$

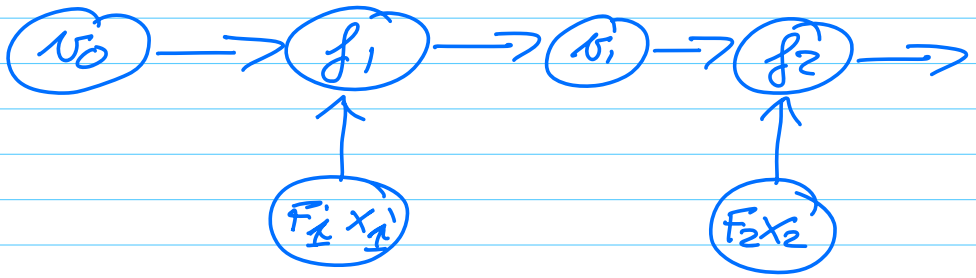
$\Delta t = \frac{t_m - t_0}{\Delta t}$

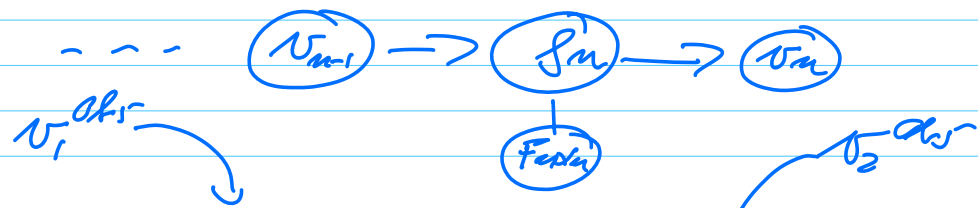
$x_{i+1} = x_i' + \Delta t v_i'$

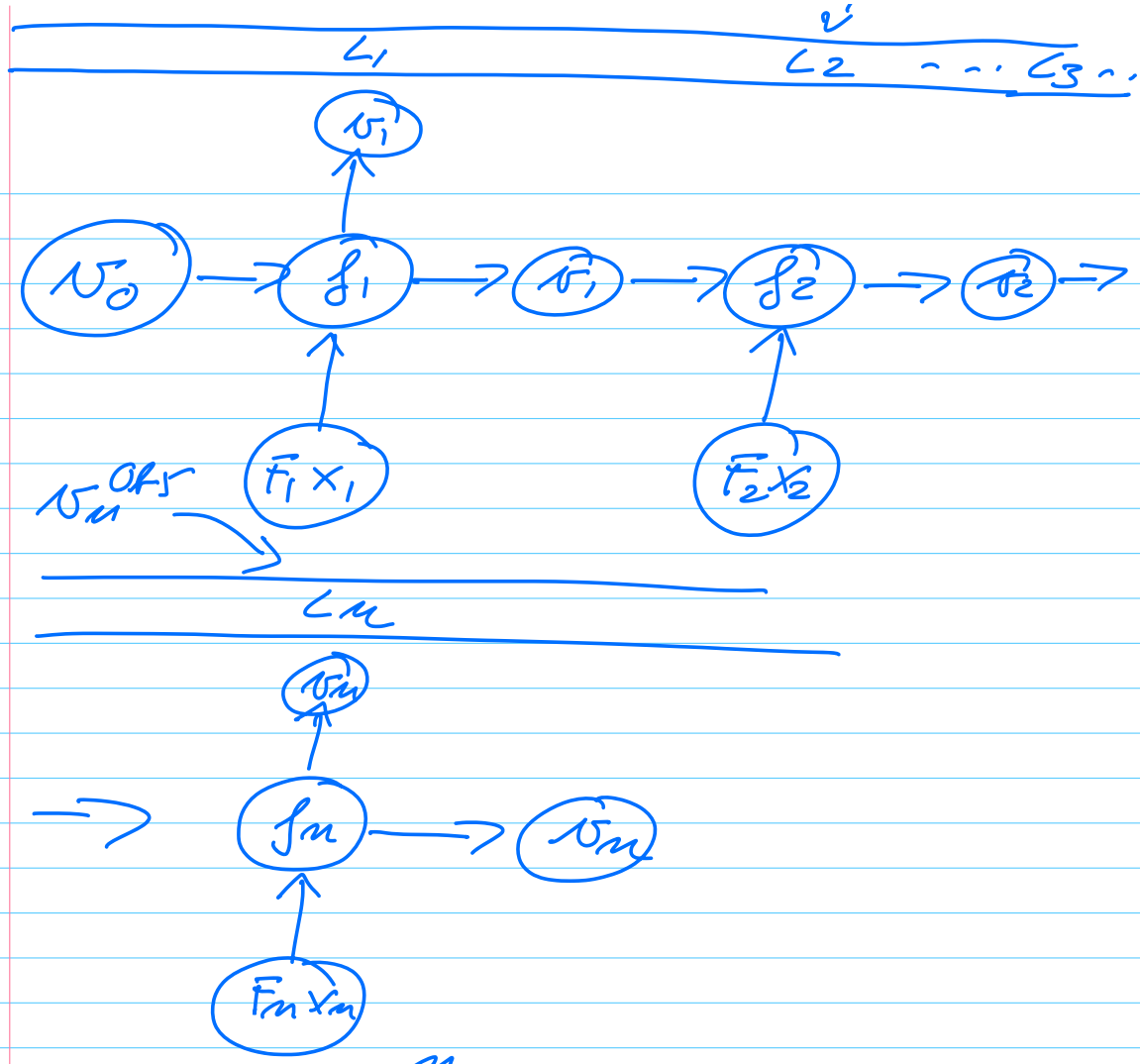
$v_{i+1} = v_i' + \Delta t f_i'$

$f_i' = f(F_i', x_i', v_i')$

Graphical representation



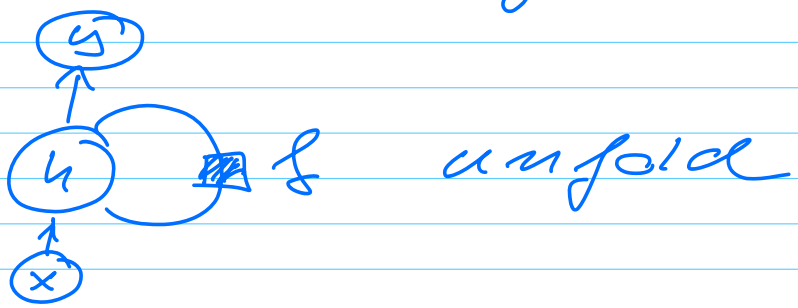


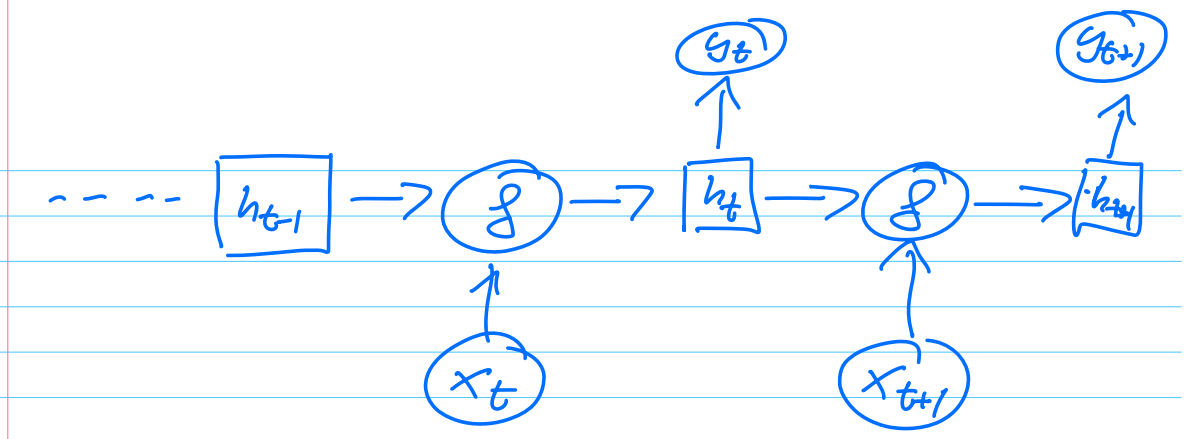


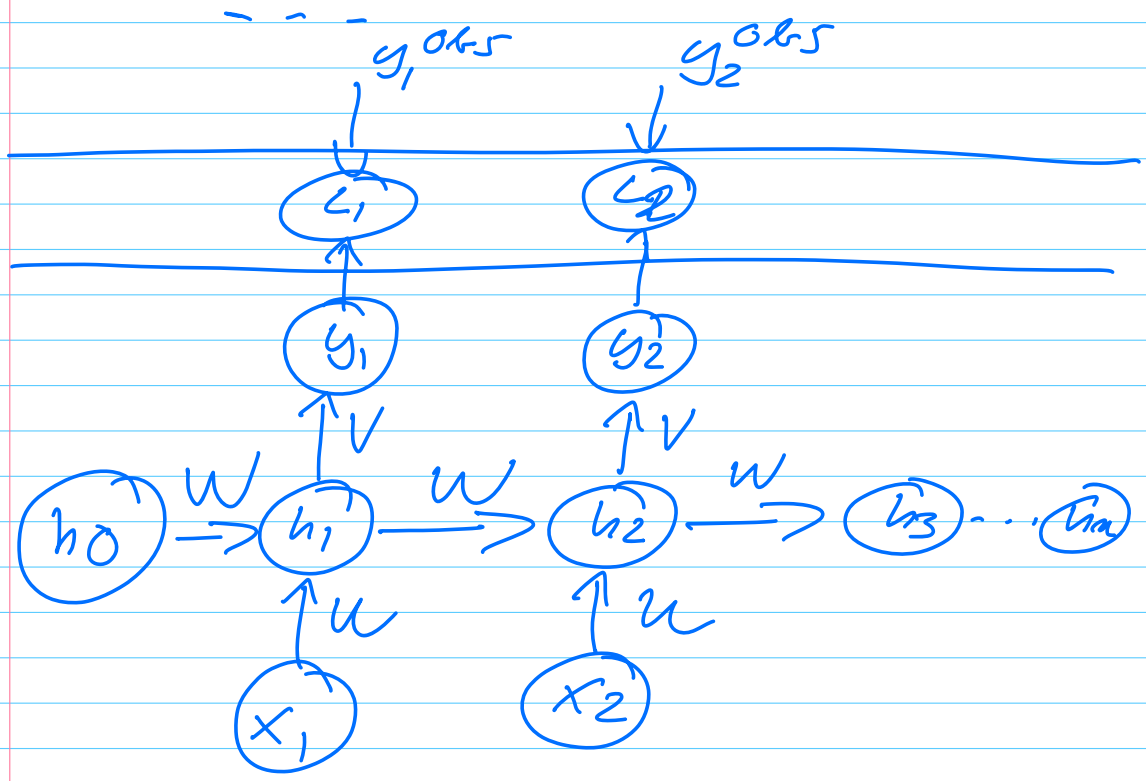
$$\mathcal{L} = \sum_{i=1}^{n} \mathcal{L}_i$$

RNN

$$N_\theta \rightarrow h_t = f(h_{t-1}, x_t; \theta)$$







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

Forward pass

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

$$h_t = f(z_t)$$

$$o_t = Vh_t + C \quad \leftarrow \text{bias}$$

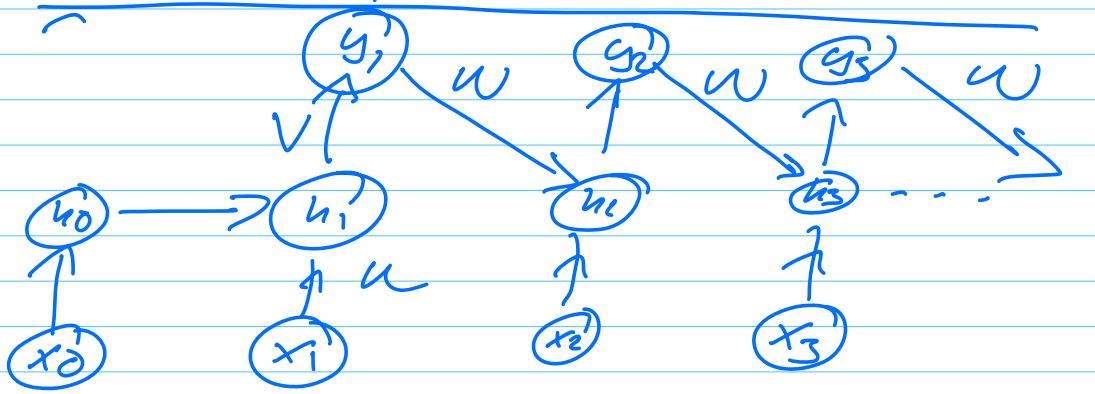
$$y_t = g(o_t)$$

Common strategies
- weights are shared
- change weights at every stage,

$$\mathcal{L} = \sum_{i=1}^{n} \mathcal{L}_i$$

Expensive to train and impossible to parallelize.

Simpler NN ← gates



Previous (and this) training is done by Back propagation through time (BPTT)

Gradients

$$\frac{\partial L_t}{\partial W} = \frac{\partial L_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \times \sum_{k=1}^{m} \left(\frac{\partial h_t}{\partial h_k} \right) \frac{\partial h_k}{\partial W}$$

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

This can lead to

- vanishing gradients
- exploding gradients

Exploding gradients (skip x_t)

$$h_t = W h_{t-1}$$

weights are reused

$$h_t = \underbrace{W \cdot W \cdots W}_{t\text{-times}} h_0$$

$$= W^t \cdot h_0$$

assume W is diagonalizable

$$W = U D U^T \qquad U U^T = U^T U = \mathbb{1}$$

expand h_0 in eigenvectors of W, w_i, with eigenvalues λ_i $W \in \mathbb{C}^{m \times m}$

$$h_0 = \sum_{i=0}^{m-1} \lambda_i w_i$$

$$W w_i = \lambda_i w_i$$

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

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

$$\lambda_0 > \lambda_1 > \lambda_2 > \dots > \lambda_{m-1}$$

$t \Rightarrow \infty$

$$W^t h_0 \simeq \lambda_0^t \alpha_0 w_0$$

$\lambda_0 > 1$ risk of exploding gradient

$\lambda_0 < 1$ risk of vanishing

gradients

exploding gradients : gradient clipping. gradient $\vec{g}$

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

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

endif.