FYS5429/9429, lecture Feb 27, 2025

FYS 5429/9429, Feb 27

Building our view of an RNN

ODE

$$\underbrace{m\frac{d^2x}{dt^2}}_{\text{acceleration}} + \underbrace{\eta\frac{dx}{dt}}_{\text{damping}} + x(t) = \underbrace{F(t)}_{\text{applied external force}}$$

prop with $v = \frac{dx}{dt}$

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

$$m\frac{dv}{dt} + \eta v + x = F$$

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

$$\frac{dv}{dt} = -\frac{\eta}{m}v - \frac{x}{m} + \frac{F}{m}$$

$$m = 1.$$

Discretize : $v \Rightarrow v_i = v(t_i)$

$x \Rightarrow x_i = x(t_i)$

$$t_i = t_0 + \Delta t \cdot i \qquad i = 0, 1, \dots, n$$

$$\Delta t = \frac{t_n - t_0}{n}$$

Euler's method

$$x_{i+1} = x_i + \Delta t \cdot v_i$$

$$v_{i+1} = v_i + \boxed{\Delta t \left(F_i - \eta v_i - x_i \right)}$$

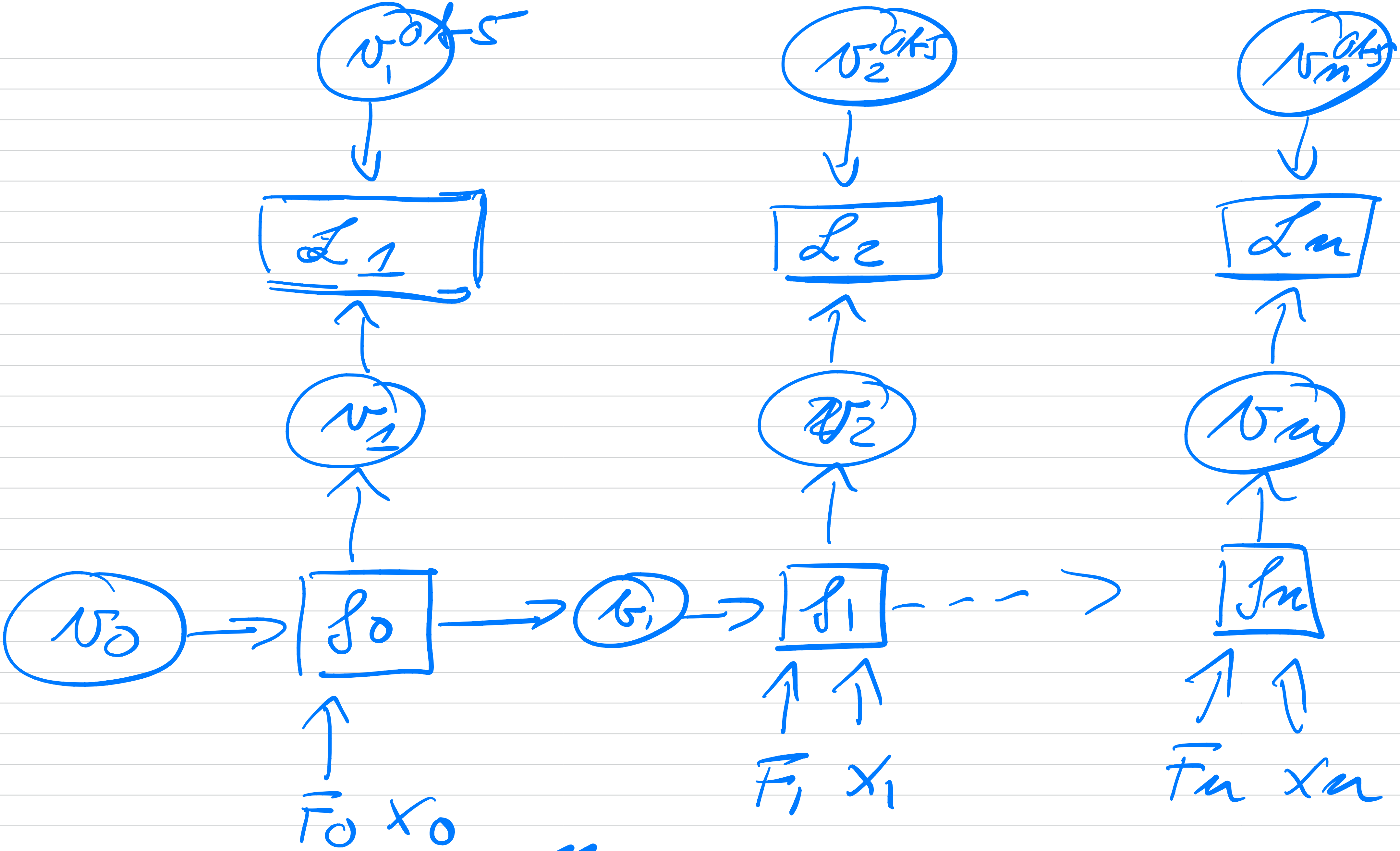
Look at v only

$$v_{i+1} = v_i + f(v_i, x_i, \Delta t, F_i)$$

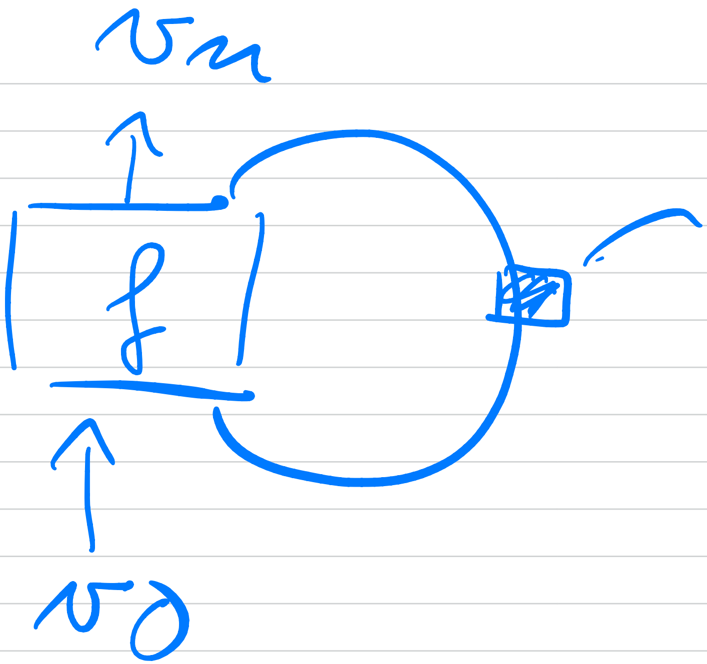
Graphical representation

$$\begin{tikzcd}
v_0 \arrow[r] & \boxed{f_0} \arrow[r] & v_1 \arrow[r] & \boxed{f_1} \arrow[r] & v_2 \cdots \\
 & F_0 \; x_0 \arrow[u] & & F_1, x_1 \arrow[u] &
\end{tikzcd}$$

$$\begin{tikzcd}
\cdots & f_n \arrow[r] & v_n \\
 & F_n \; x_n \arrow[u] &
\end{tikzcd}$$



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



different operations

set up a simple RNN

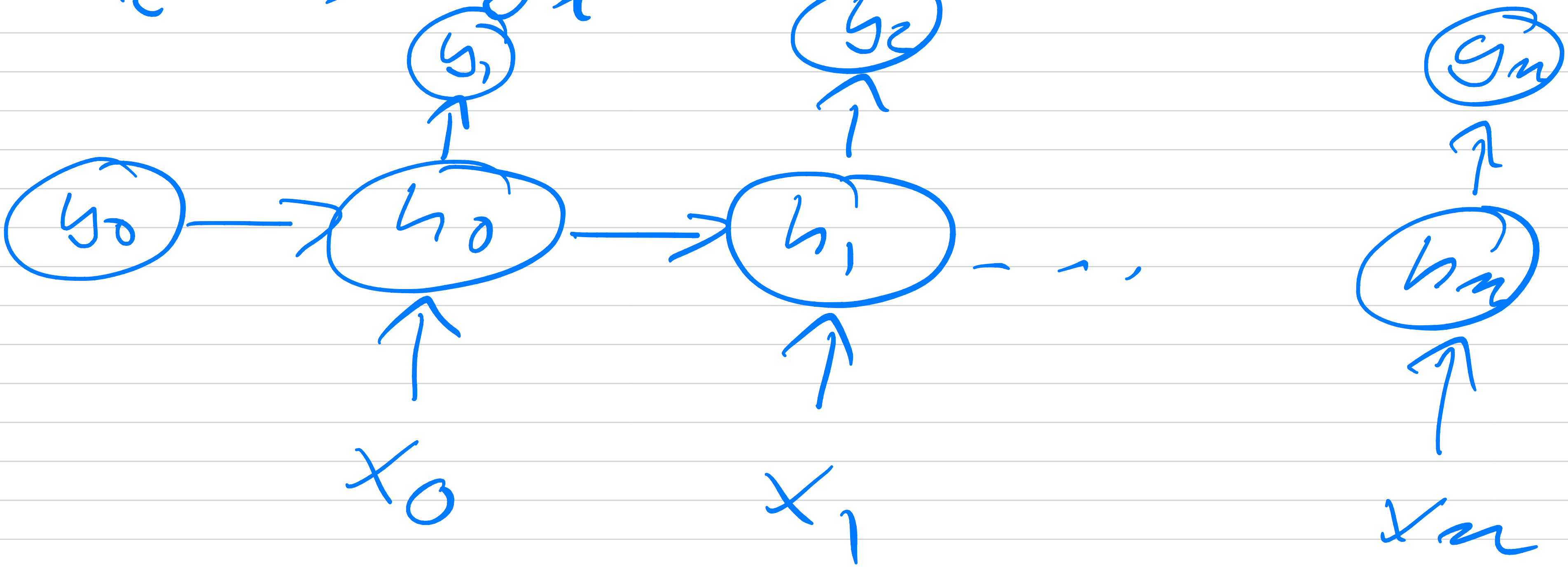
— skip F_i'

$$v_{i+1} = \underbrace{v_i + f(F_i', v_{i-1}, x_i' \ldots)}$$

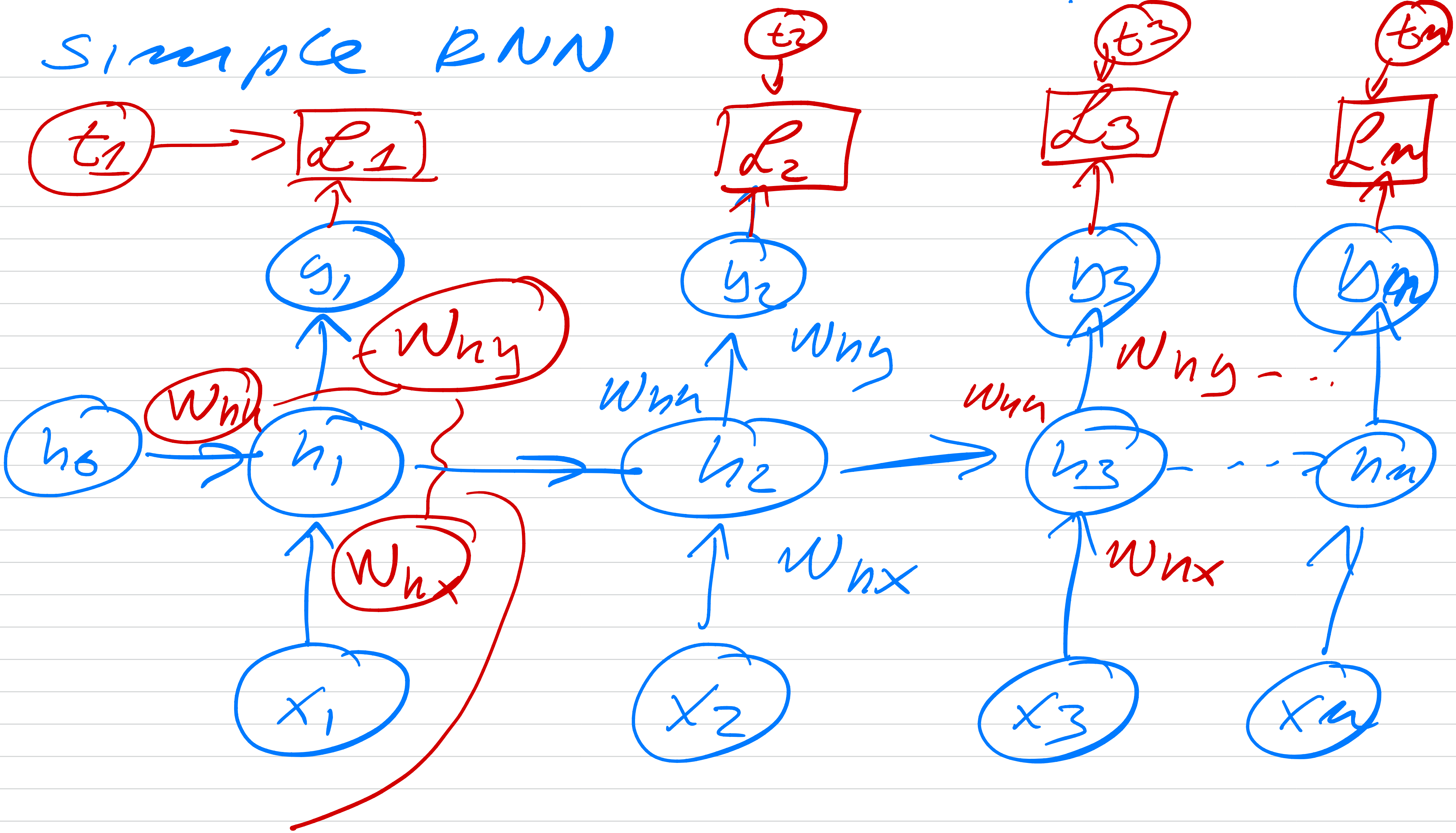
$$h(x_i', (v_i = h_{i-1})) = h_i'$$

$$v_{i+1} = h_i' \quad , \quad v_i = h_{i-1}'$$

$v_i' \longrightarrow y_i'$



Simple RNN



Parameters to be trained

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \mathcal{L}_i(\theta, t_i, y_i)$$

$$h_i' = \sigma(z_i')$$

activation function

$$z_i' = W_{hx} x_i' + W_{hh} h_{i-1}' + b_i'$$

bias

$$y_i' = \sigma_y(a_i')$$

$$a_i' = W_{hy} h_i' + c_i'$$ — bias

$$\Theta = \{ W_{hh}, W_{hy}, W_{hx}, b, c \}$$

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

- Training is done iteratively
  - Feed Forward pass
  - Back propagation in time

$\nabla_c \mathcal{L}$, $\nabla_b \mathcal{L}$, $\nabla_{W_{hh}} \mathcal{L}$, $\nabla_{W_{hx}} \mathcal{L}$

$\nabla_{W_{hy}} \mathcal{L}$

can lead to exploding or vanishing

Exploding/vanishing gradients

problems with W_{hh} and

derivatives like

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

which is a part of $\frac{\partial h_i}{\partial W_{hh}}$

assume weight sharing

$W_{hh} = W$ (leave out bias)

$h_i = W h_{i-1}$

$$h_i = \underbrace{W \cdot W \cdots W}_{i} h_0$$

$$= W^i h_0$$

assume W can be diagonalized

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

eigenvalues

$$D = \begin{bmatrix} \lambda_0 & & & \\ & \lambda_1 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

$$W \in \mathbb{R}^{m+1 \times m+1}$$

$$W w_i' = \lambda_i' w_i'$$

$$h_0 = \sum_j \alpha_j' w_j'$$

$$W h_0 = \sum_j \alpha_j' W w_j' = \sum_j \alpha_j \lambda_j w_j$$

Repeat t-times

$$W^t h_0 = h_t = \sum_j \lambda_j^t \alpha_j w_j$$

$$\lambda_0 \geq \lambda_1 \geq \lambda_2 \ldots \geq \lambda_m$$

$\lim_{t \to \infty} h_t \approx \lambda_0^t \alpha_0 w_0$

if $\lambda_0 > 1$, then gradients can explode

if $\lambda_0 < 1$, then gradients can vanish

$\vec{g}$ = gradient, gradient clipping

if $\quad \|\vec{g}\|_2 \geq \varepsilon$ (fixed number)

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

endif