FYS5429/9429 February 27

FYS5429/9429 February 27

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

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

$$\frac{dv}{dt} = -\underbrace{\frac{\eta}{m}}_{\alpha} v - \underbrace{\frac{x}{m}}_{\delta} + \underbrace{\frac{F}{m}}_{\tilde{F}}$$

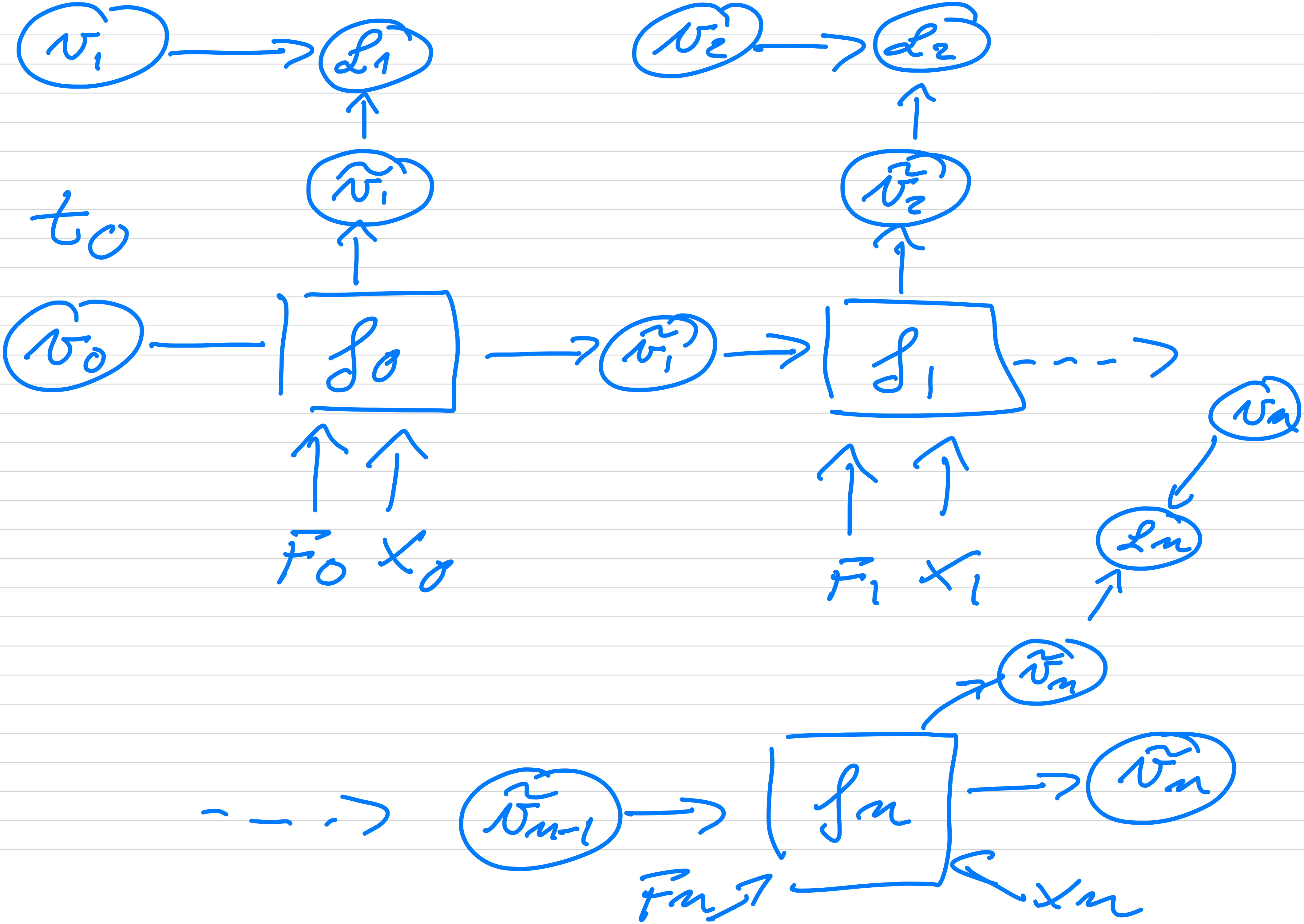
$$v_{i+1} = v_i + \Delta t\left(\tilde{F}_i - \alpha v_i - \delta x_i\right)$$

$$v_i + f(v_i, t_i, \tilde{F}_i, \Delta t)$$

$t_0 \qquad v(t_0) = v_0$

$$\begin{tikzcd}
v_0 \arrow[r] & f_0 \arrow[r] & v_1 \arrow[r] & f_1 \arrow[r] & v_2 \\
 & F_0 \; X_0 \arrow[u] & & F_1 \; X_1 \arrow[u] &
\end{tikzcd}$$

$$\begin{tikzcd}
{--} \arrow[r] & {} & v_{n-1} \arrow[r] & f_N \arrow[r] & v_n \\
 & & F_n \arrow[ur] & X_n \arrow[u] &
\end{tikzcd}$$


v_1
$\mathcal{L}_1$
v_2
$\mathcal{L}_2$
$\tilde{v}_1$
$\tilde{v}_2$
t_0
v_0
f_θ
$\tilde{v}_1$
f_1
v_n
F_0
x_θ
F_1
x_1
$\mathcal{L}_n$
$\tilde{v}_n$
$\tilde{v}_{n-1}$
f_n
$\tilde{v}_n$
F_n
x_n

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

$$\hat{\theta} = \underset{\theta}{\arg\min}\ \mathcal{L}(\theta)$$

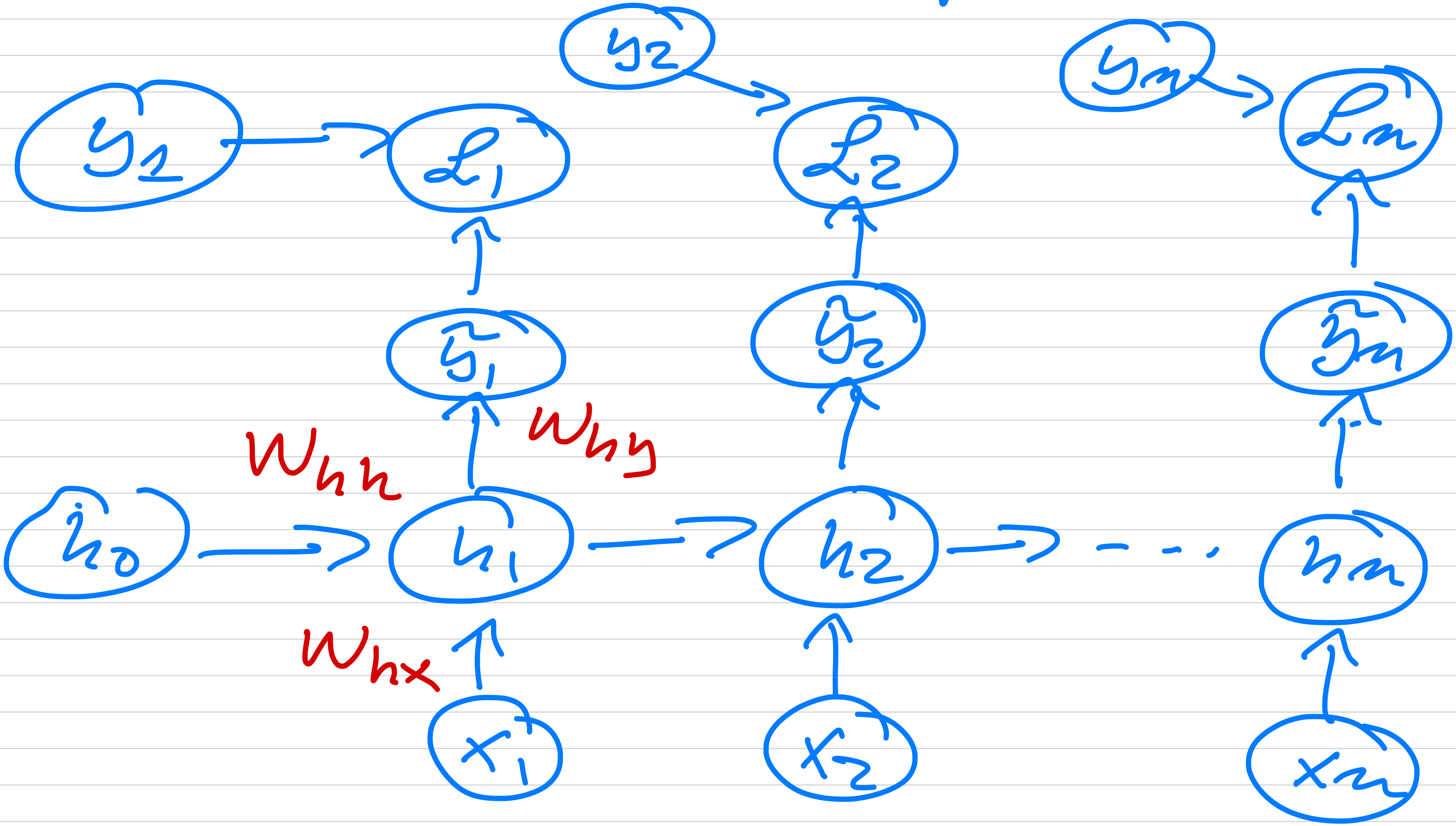
$$v_{i-1} = h_{i-1}$$

$$v_{i+1} = h(x_i, h_{i-1}, F_i)$$
$$= h_i$$

$$v = \{v_0, v_1 \ldots, v_n\}$$

$y_i \rightarrow y_1$ (output at step t_1)

$\tilde{y}_1$ = model output at t_1



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

$$\tilde{y}_i = W_{hy} \cdot h_i + c_i$$

$$\mathcal{L}(\theta) = \sum_{i=0}^{final} \mathcal{L}_i$$

$$\mathcal{L}_i(\tilde{y}_i, y_i)$$

Backpropagation in time

- Feed Forward pass
- Backprop 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 give rise to exploding

$W_{hh} \rightarrow W \qquad \tilde{y}_i \rightarrow y_i$

$$\frac{\partial L_i}{\partial W} = \frac{\partial L_i}{\partial y_i} \frac{\partial y_i}{\partial h_i} \frac{\partial h_i}{\partial W}$$

$$\frac{\partial h_i}{\partial W} = \sum_{k=0}^{n-1} \frac{\partial h_i}{\partial h_k} \frac{\partial h_k}{\partial W}$$

Computed as a multiplication of adjacent "time" step

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

W is the same for all "time" steps (weight sharing)

$$h_i = W \cdot h_{i-1}$$

$$h_i = W \cdot W \cdot W \cdots W \cdot h_0$$

$$= W^i h_0$$

Suppose we can diagonalize W

$$W w_i = \lambda_i w_i$$

$$h_0 = \sum_i \alpha_i w_i$$

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

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

repeat $-t-$ times

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

$$= \sum \alpha_i \lambda_i^t w_i$$

This sum with λ_i^t can vanish or explode or stabilize
Depends on the magnitude of λ_i

$$\lambda_0 > \lambda_1 > \lambda_2 \ldots > \lambda_m$$

$$\lim_{t \to \infty} w^t h_0 \simeq \lambda_0^t \alpha_0 w_0$$

$\lambda_0 > 1$, then the terms grows quickly $\Rightarrow$ exploding gradients gradient clipping

$\vec{g}$ = gradient

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

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

endif