FYS-STK3155/4155, OCT 28, 2022

Recurrent NN:

Example: ODE

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

initial conditions

$$x_0 = x(t_0) \wedge v_0 = \cancel{x}\, v(t_0)$$

Euler's method:

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

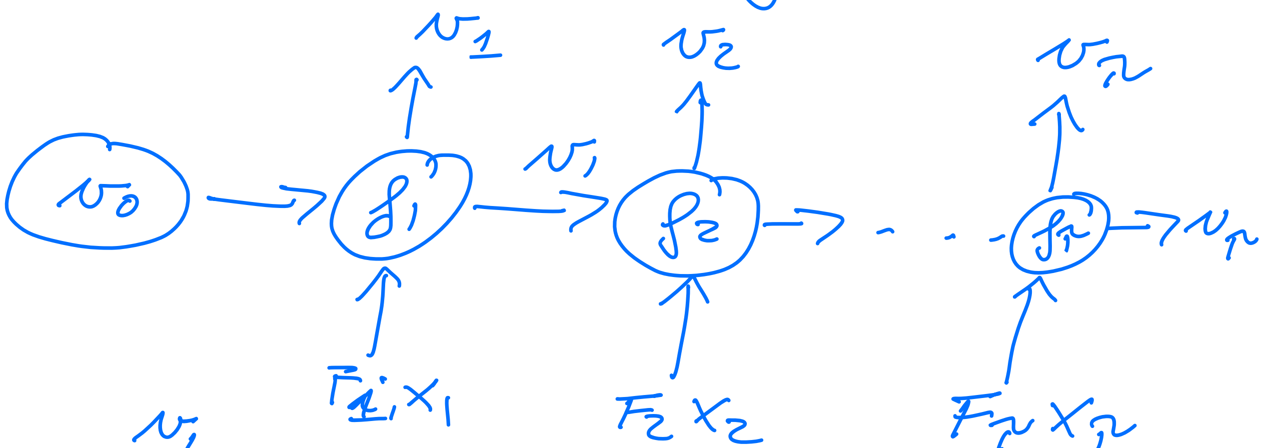
$$v = \frac{dx}{dt}$$

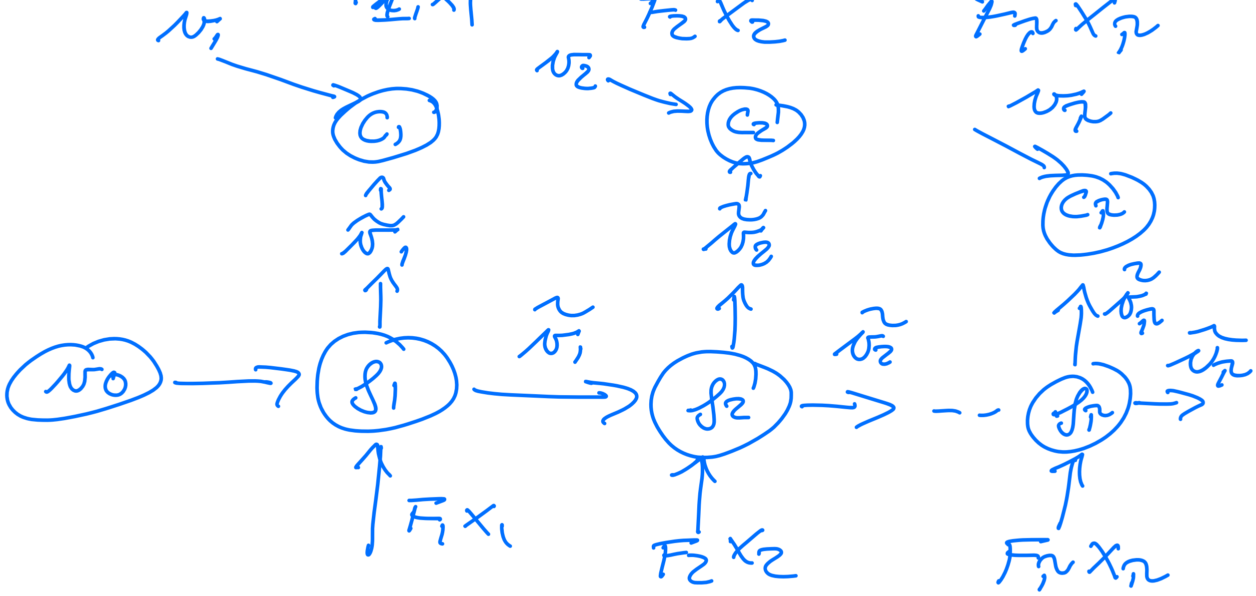
$$\frac{dv}{dt} = -\frac{\eta}{m}v - \frac{1}{m}x + \frac{F}{m} = \tilde{F} - \alpha v - \delta x$$

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

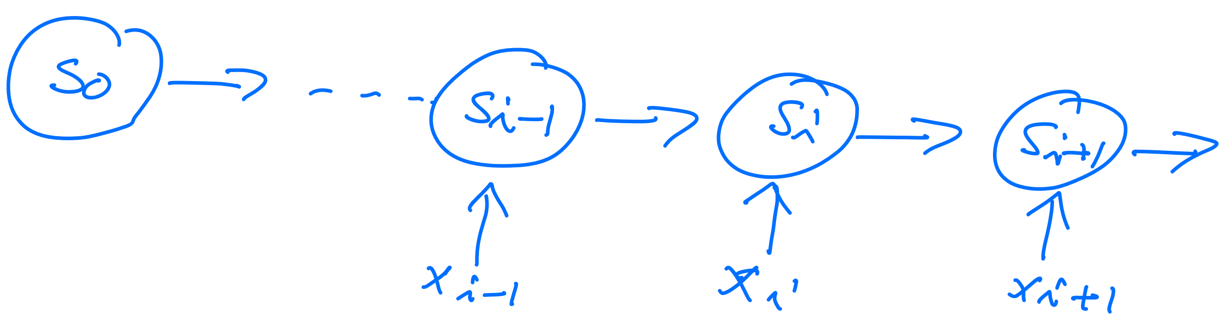
$$= f(v_i, \tilde{F}_i, x_i)$$

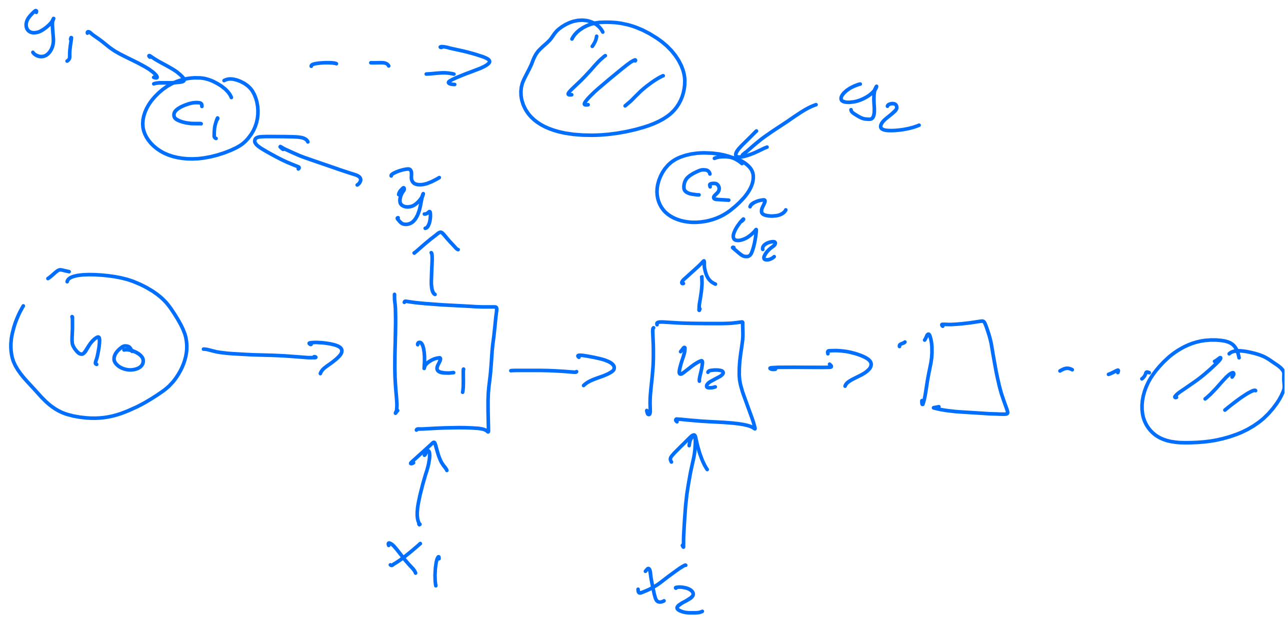
Look at v_i only



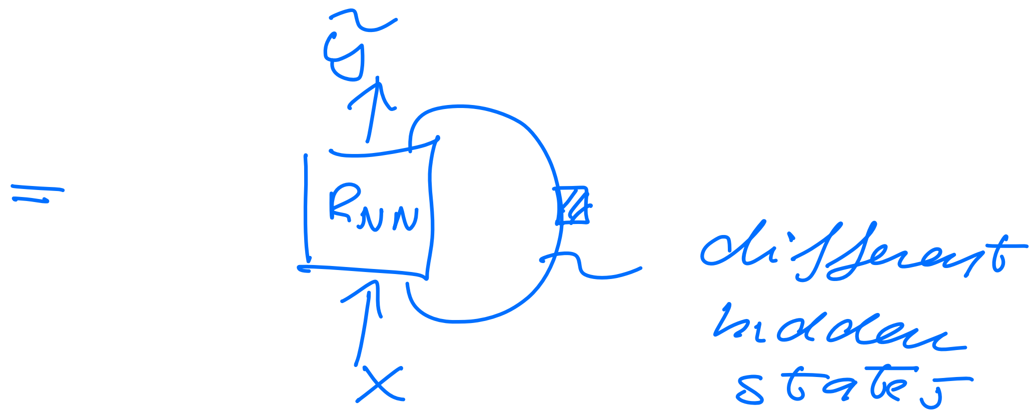


$$v_{i+1} \Rightarrow s_{i+1} = h(s_i, x_i; \theta)$$
$$= h_i$$

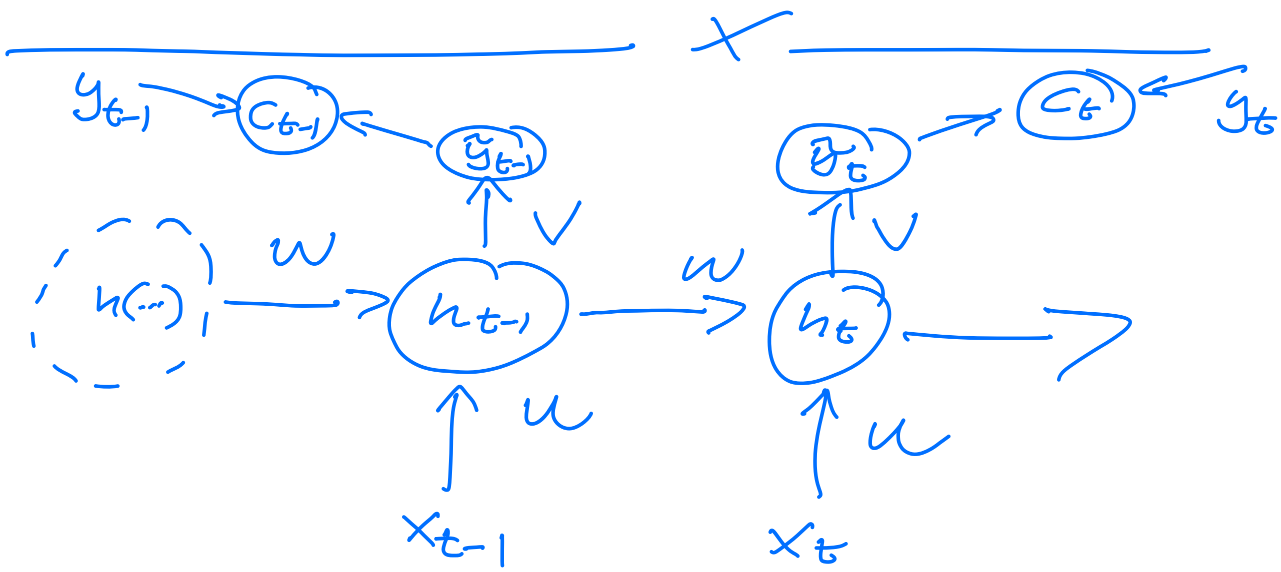




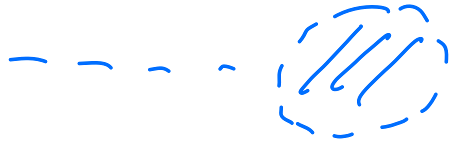
y1
c1
y2
c2
h0
h1
h2
x1
x2



=
RNN
x
different hidden states



yt-1
Ct-1
yt-1
Ct
yt
h(...)
W
ht-1
V
W
ht
V
U
U
xt-1
xt



bias

$$z_t = b + Wh_{t-1} + Ux_t$$

$$h_t = \sigma(z_t)$$

bias

$$\tilde{y}_t = g(h_t \cdot V + c)$$

The cost function at each time step

$$C_t = C_t(x_t, h_t, \tilde{y}_t, y_t)$$

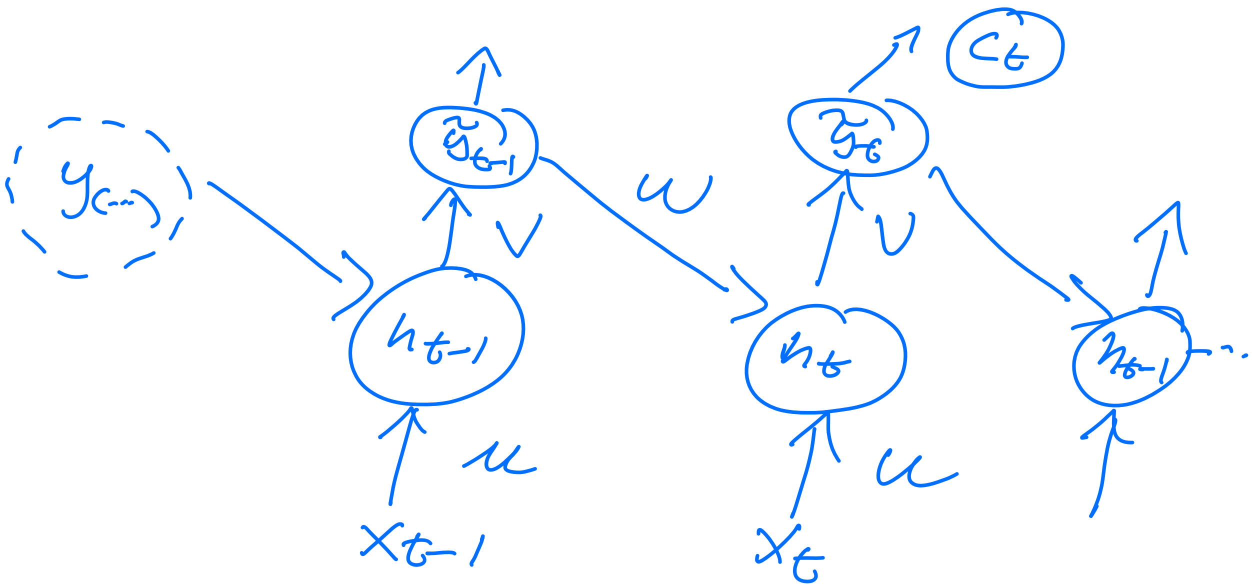
BPTT = Back propagation through time.

Reduction of computational complexity.

y_{t-1}

C_{t-1}

y_t



Problem with RNNs are often due to exploding gradients

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

This operation is repeated for h_t t-times

$$h_t = (W)^t h_0$$

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

$$D = \begin{bmatrix} \lambda_0 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix}$$

eigenvalues of W are λ_i

eigenvectors w_i

$$h_0 = \sum_i \alpha_i w_i$$

$$W h_0 = h_1 = \sum_i \alpha \lambda_i w_i$$

$$W w_i = \lambda_i w_i$$

Repeat t-times

$$(W)^t h_0 = \sum_i \alpha \lambda_i^t w_i$$

$$\lambda_0 > \lambda_1 > \lambda_2 > \dots > \lambda_d$$

when t is large

$$h_t \simeq \lambda_0^t w_0 \alpha_0$$

if $\lambda_0 \geq 1$, we may get contributions to h_t which become very large $\Rightarrow$ can give rise to exploding gradients.

To avoid this it is common to use gradient clipping

— $\vec{g}$

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

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

end if