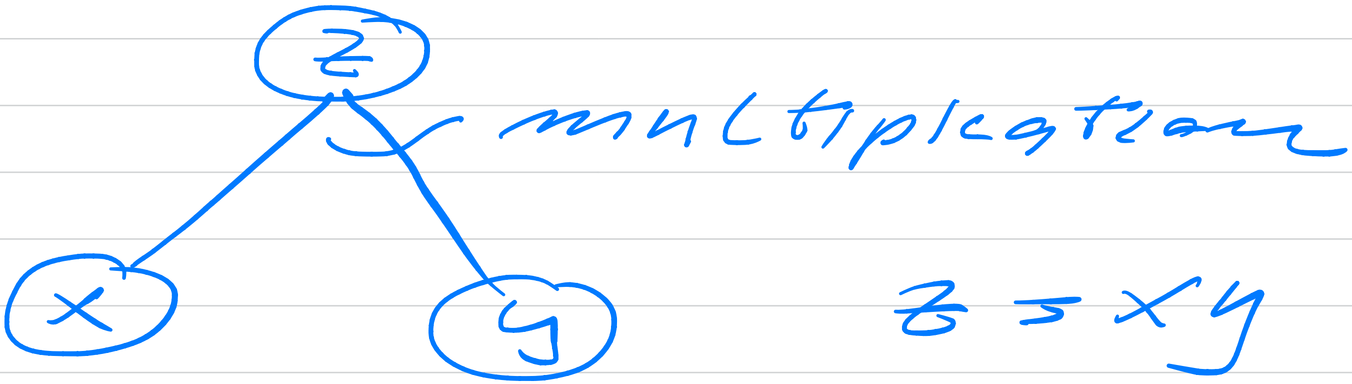
Lecture Fys-Stk3155/4155, November 9, 2023

Recurrent NNs

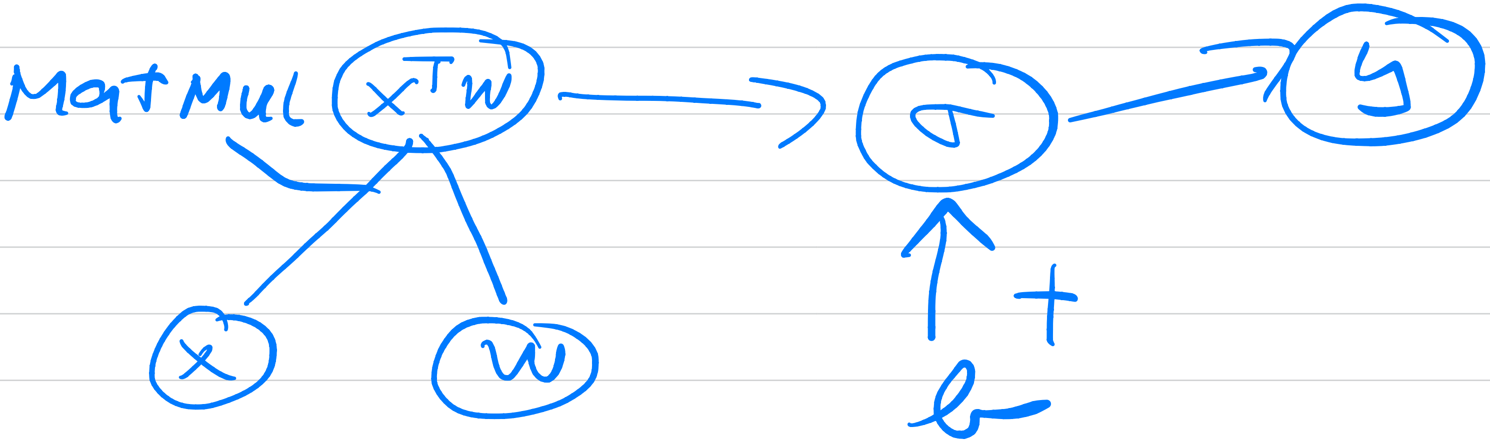
Graphical representation



$$z = xy$$

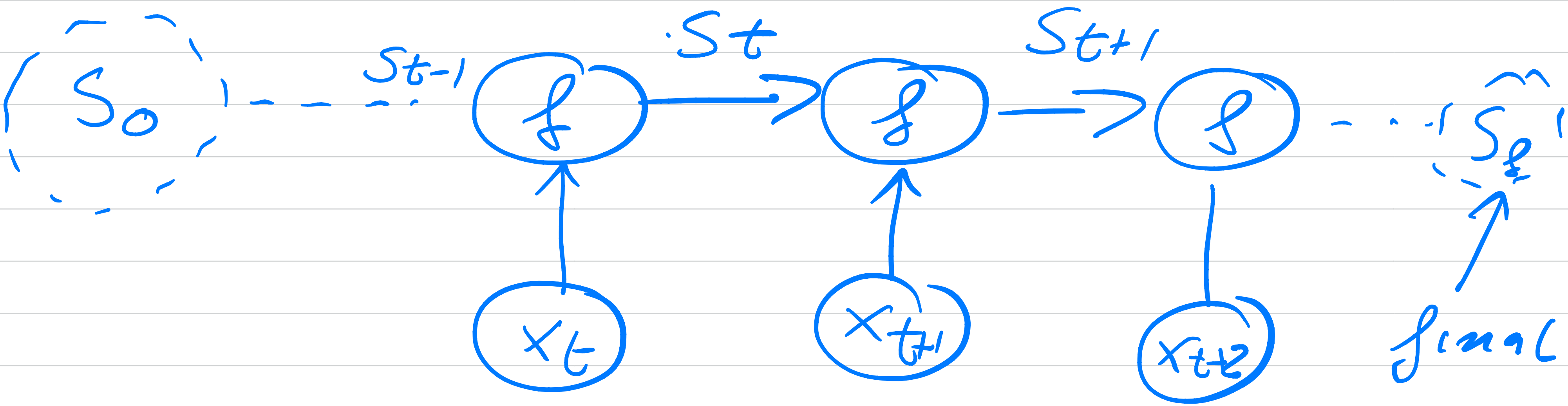
NN:

$$y = \sigma(x^T w + b) = \sigma(z)$$



Think of a dynamical system driven by an external signal at a time $-t-$

$$S_t = f(S_{t-1}, x_t; \theta)$$



Example

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

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

rewrite as two coupled differential (Focus on v)

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

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

$$= f(F, x, v)$$

$x \to x_i \qquad i = 0, 1, 2, \ldots n$

$v \to v_i \qquad \Delta t = \dfrac{t_f - t_0}{n}$

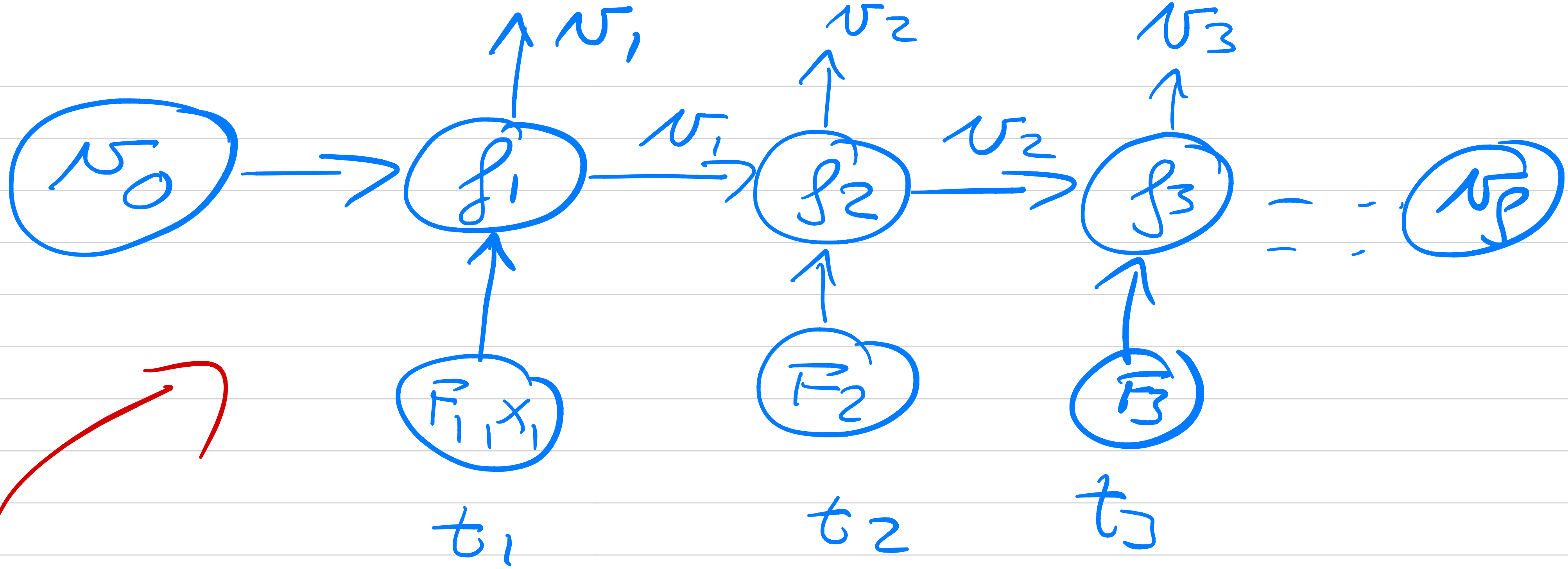
$t_i = t_0 + i\Delta t$

Euler's method

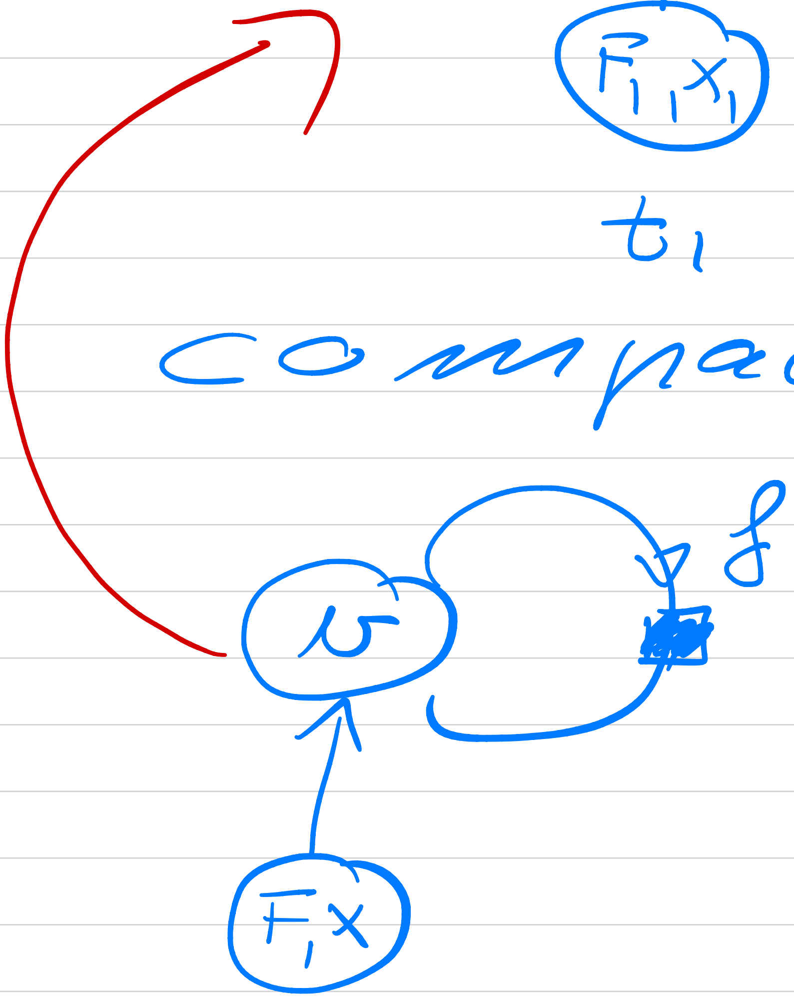
$x_{i+1} = x_i + \Delta t \cdot v_i$

$v_{i+1} = v_i + \Delta t \cdot f_i$

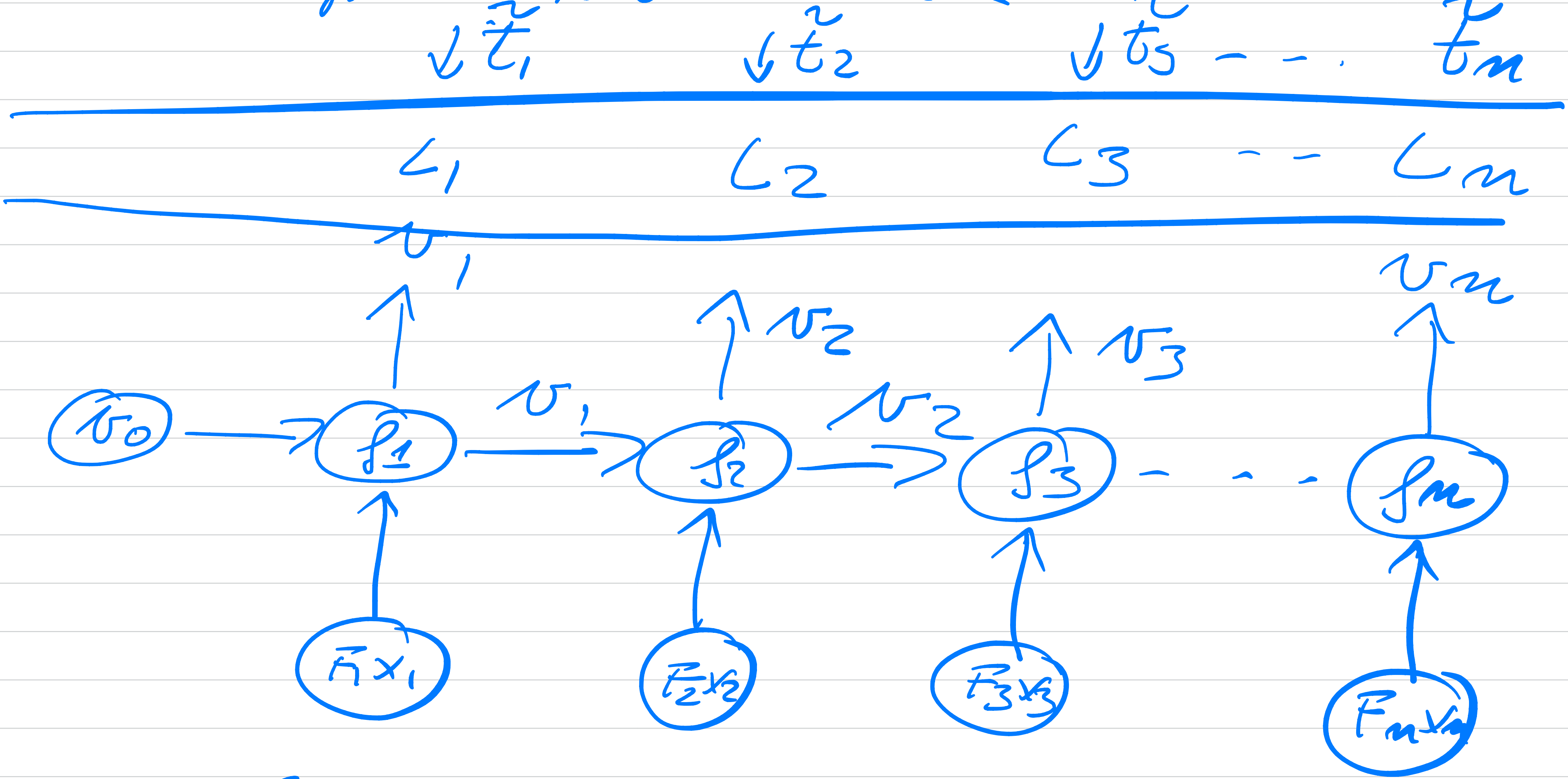
$f_i = f(F_i, x_i, v_i)$



compact rewrite



Simple Network

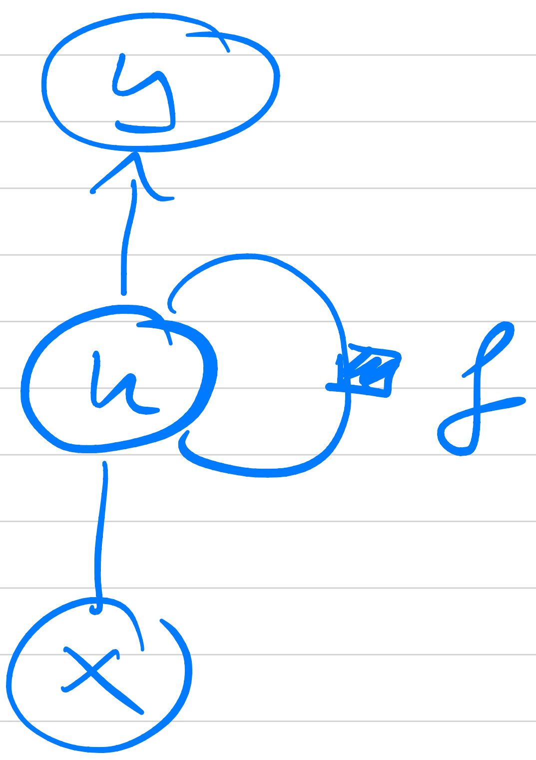


$\tilde{t}_i$ = observation at time t_i (target, output)

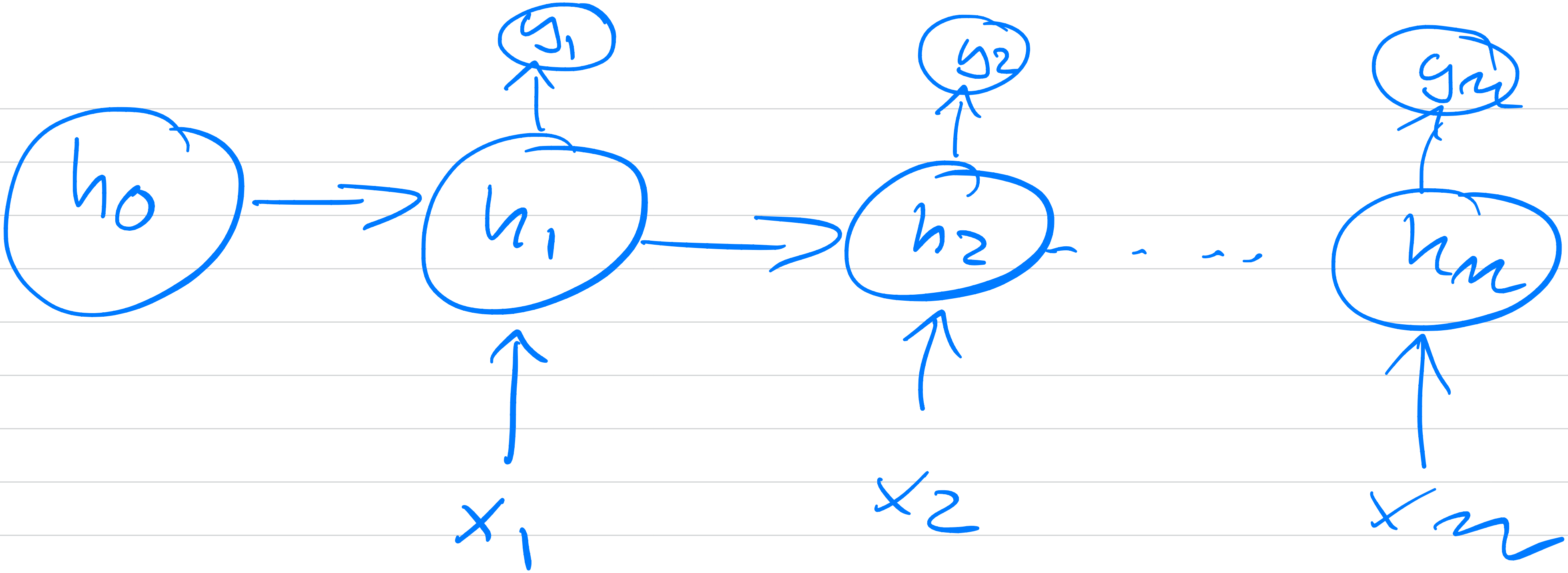
Now we replace the function -f- with a neural net

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

set of parameters to train

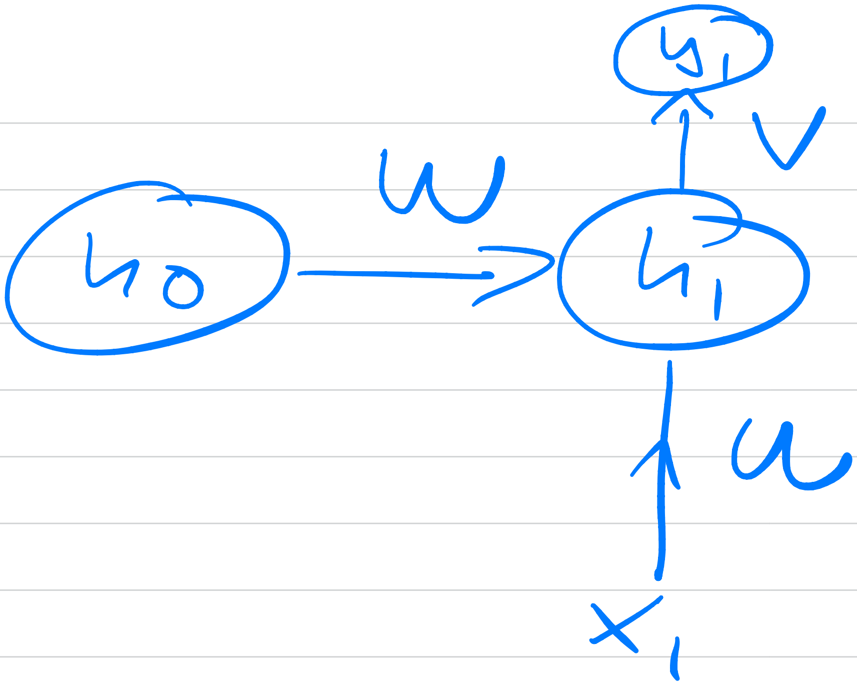


open up



h_i refers to hidden parts

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

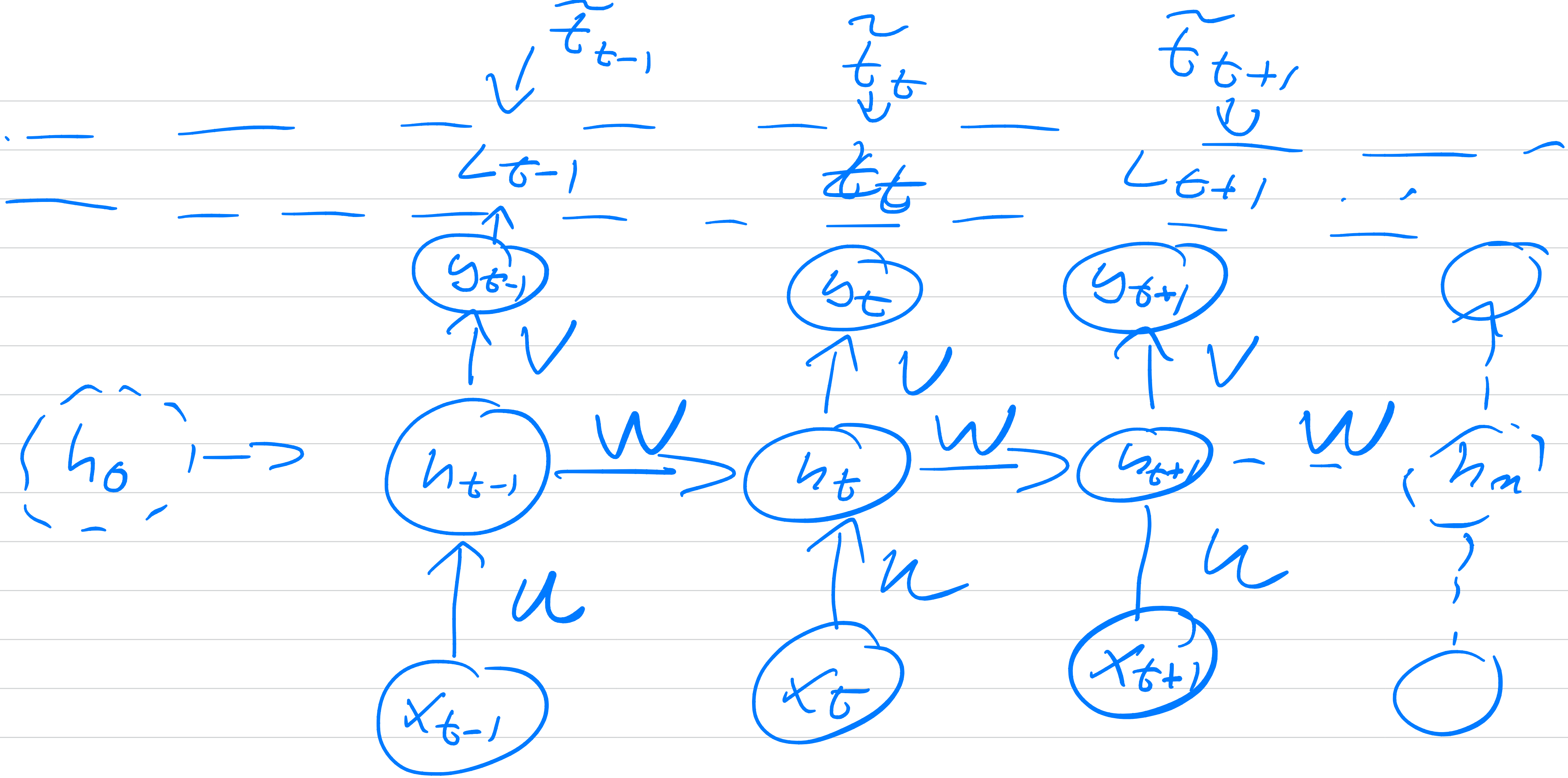


$$z_t = U x_t + W h_{t-1} + b$$ ← bias

$$h_t = f(z_t)$$ (often tanh or ReLU)

$$r_t = V h_t + c$$ ← bias

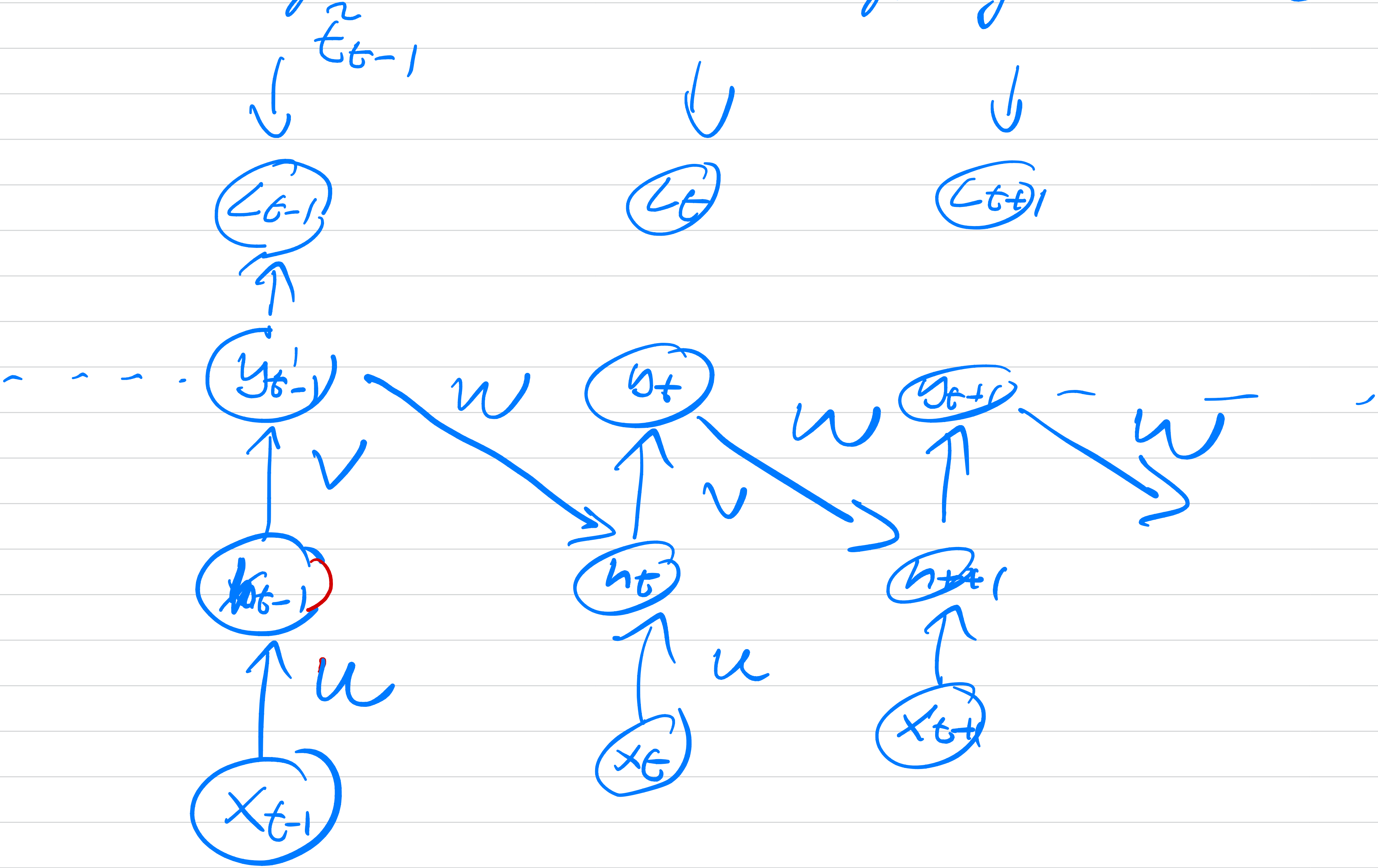
$$y(t) = g(r_t)$$



$$\mathcal{L} = \sum_{i=0}^{n} \mathcal{L}_i$$

Back propagation in time (BPT) leads to a sequential

One possible simplification



For each node (at time-t)

$\nabla_u \mathcal{L}_t$, $\nabla_W \mathcal{L}_t$, $\nabla_V \mathcal{L}_t$

$\nabla_b \mathcal{L}_t$, $\nabla_c \mathcal{L}_t$

challenges is exploding or vanishing gradients

$$\frac{\partial \mathcal{L}_t}{\partial W} = \frac{\partial \mathcal{L}_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \left[\sum_{k=1}^{n} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W} \right]$$

Example

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

assume that W is the same for all previous t-

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

$$= W^t h_0$$

suppose W is diagonali-zable, $u^T W u = D$

Expand h_0 in terms of eigenvectors w_i, eigenvalues λ_i

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

$$h_1 = W h_0 = \sum_i \alpha_i W w_i$$

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

Repeat t-times

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

$\lambda_0 > \lambda_1 > \lambda_2 \dots > \lambda_n$

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

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

if $\lambda_0 < 1$, can lead to vanishing gradients.

Gradient clipping can be used to avoid exploding gradients.

gradient $\vec{g}$

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

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

endif.