Lecture october 28

RNNs Basic definitions

Example:

ODE:

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

$$v = \frac{dx}{dt} \qquad v(t_0) = v_0$$
$$x(t_0) = x_0$$

$$a = \frac{dv}{dt}$$

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

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

Euler's method (velocity)

$$v_{i+1} \approx v_i + \Delta t v_i' \Big|_{t=t_i}$$

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

$$= -\tilde{x} + \tilde{F}$$

$$v_{i+1} \simeq v_i + \Delta t \left(-\tilde{x}_i + \tilde{F}_i\right)$$

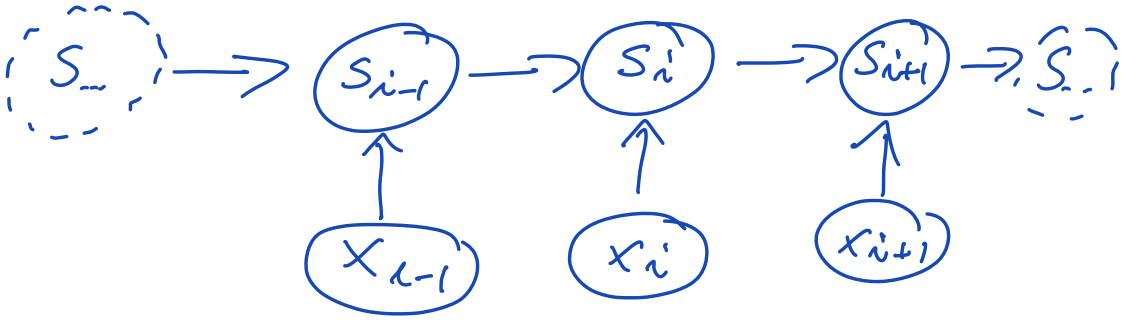
$$v_{i+1} = f\left(\tilde{x}_i, \tilde{F}_i, v_i\right)$$

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

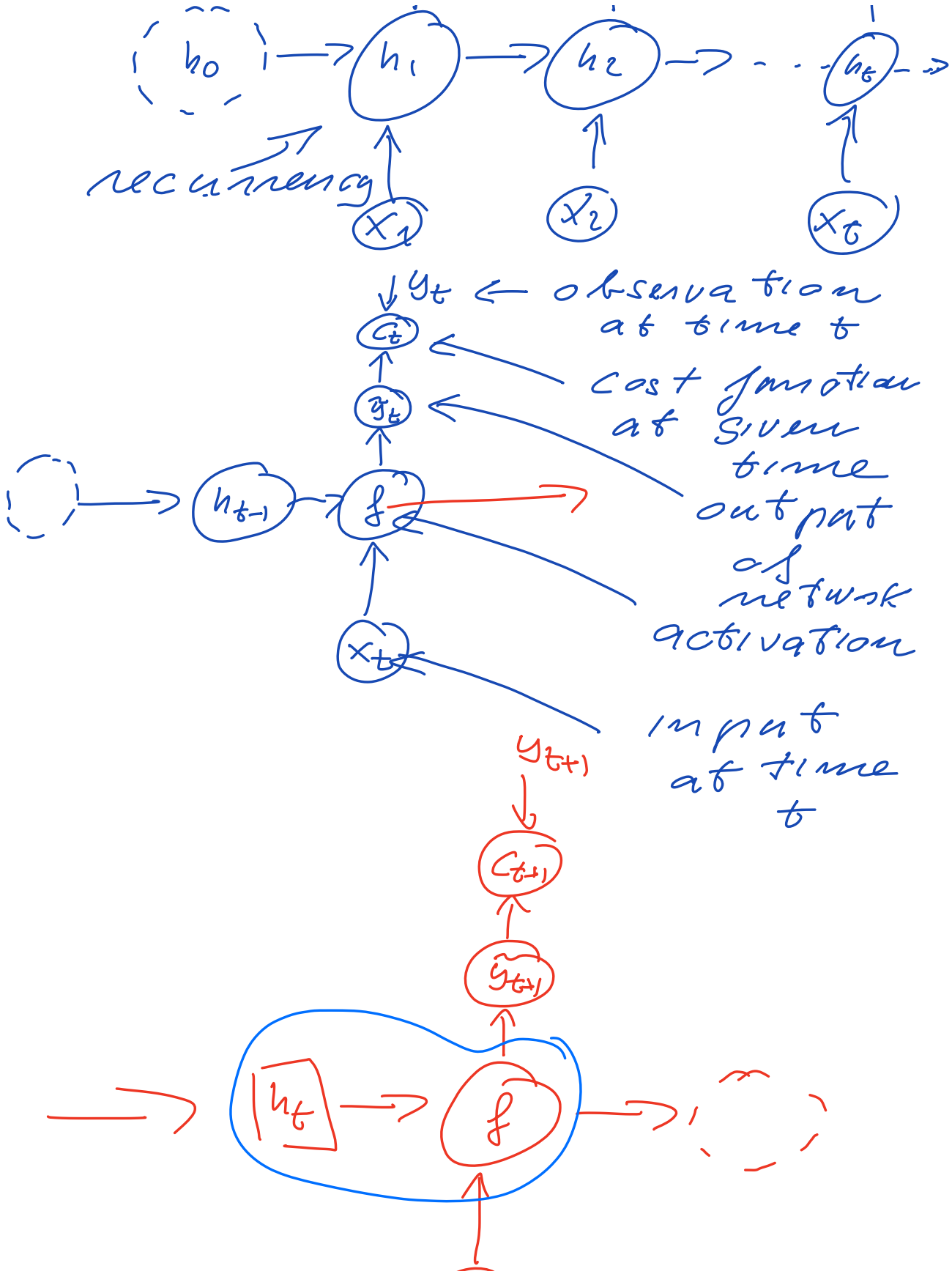
$$v_{i+1} = f\left(f(x_{i-1}, F_{i-1}, v_{i-1}), F_i, x_i\right)$$

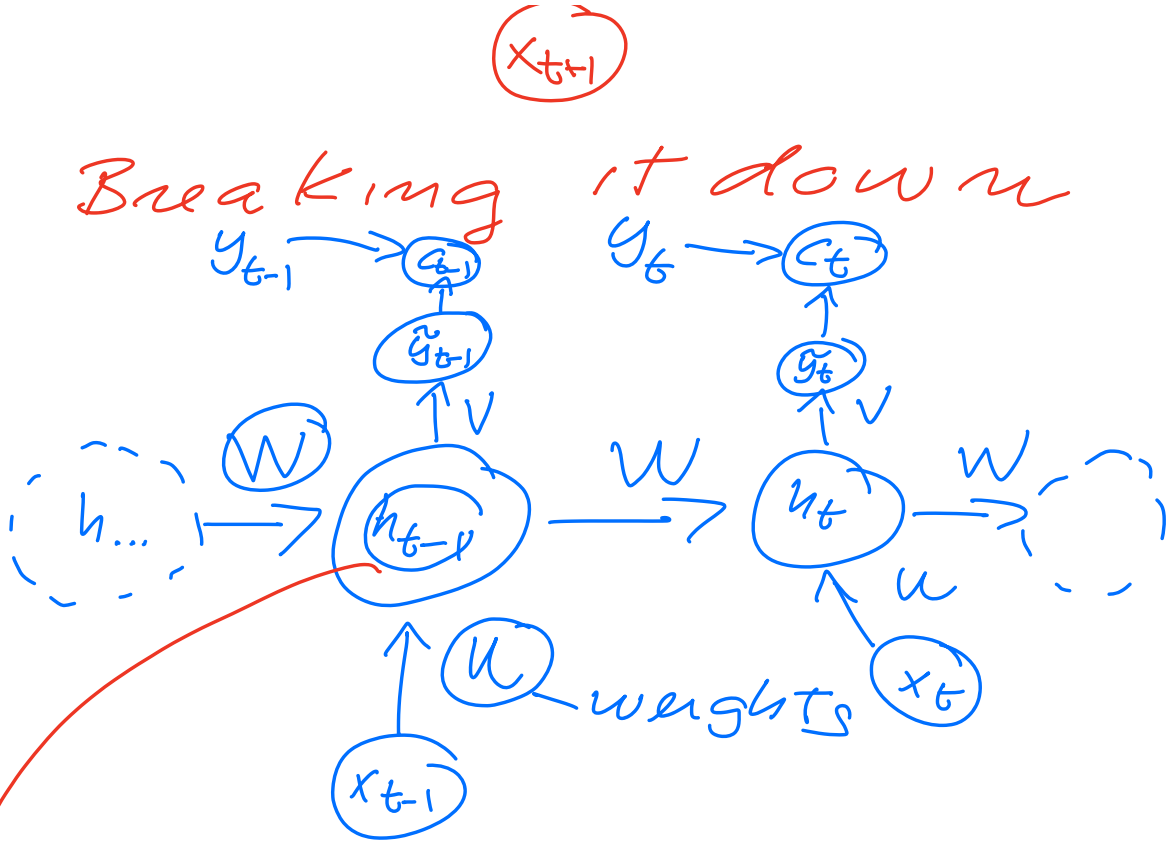
$$\downarrow$$

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



expand to include an output y_t


h_0
h_1
h_2
h_t
recurrency
x_1
x_2
x_t
y_t ← observation at time t
c_t
cost function at given time
$\tilde{y}_t$
output of network activation
h_{t-1}
f
x_t
input at time t
y_{t+1}
c_{t+1}
$\tilde{y}_{t+1}$
h_t
f



parameters:

W (weights)

$h_{t-1} = f(z_{t-1})$

$z_{t-1} = b + Wh_{t-2} + Ux_{t-1}$

$\tilde{y}_{t-1} = g(h_{t-1}V + c)$

the cost function at time $t-1$

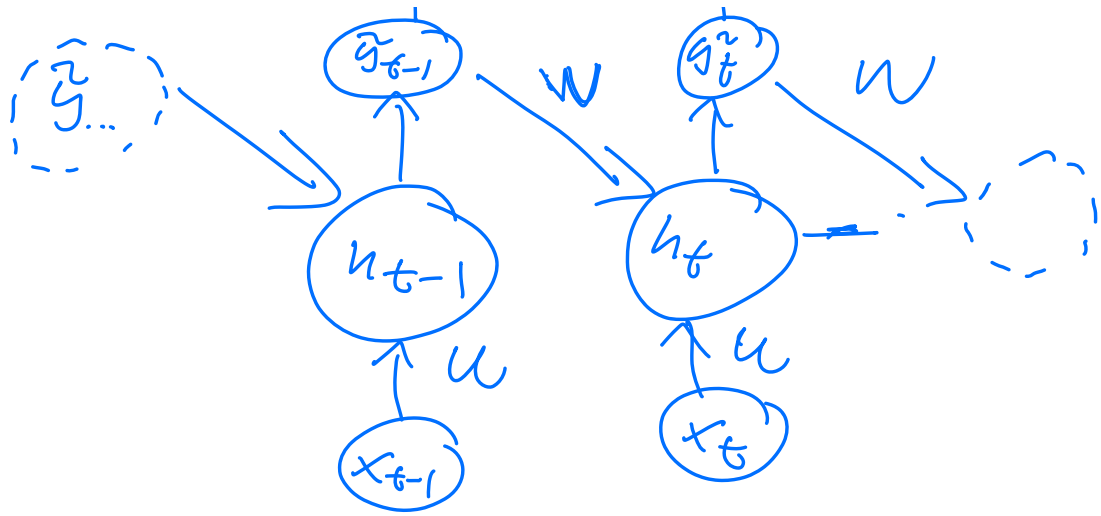
$C_{t-1}(y_{t-1}, h_{t-1}, \tilde{y}_{t-1}, x_{t-1})$

1) Feed Forward part with initialization of weights and biases at all $t-1$

2) Back propagation through time. (BPTT)

Memory and CPU intensive and keeps long term memory.

- instead of connections between hidden layers, we can feed in output $\tilde{y}_{t-1}$ into h_t



Gradient training problems

- vanishing gradients
  less of a problem
- exploding gradients

simple example

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

↑ output from hidden nodes (no x_t no bias)

W is the same at all times -t-

$h_1 = W I, W W I, \ldots W^t I$

$$= W^t h_0$$

$$W w_i = \lambda_i w_i$$

$$h_0 = \sum_i \alpha_i w_i$$

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

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

$\lambda_0 > 1$, then $W^t h_0$ will diverge,

$\Rightarrow$ ∇_{h_t} they will diverge

$\lambda_0 < 1$ the $W^t h_0$ will decrease $\Rightarrow$

$\nabla_{h_t} \Rightarrow 0$

To avoid exploding gradients :

gradient clipping

gradient g.

if $\|g\|_2 \geq \varepsilon$

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

end if.