

Lecture Twenty - Five

- ① AR & Nonlinear AR
- ② RNN
- ③ GRU
- ④ LSTM

AR & NAR

$$y_1, \dots, y_n$$

$$y_t, \quad x_t = (y_{t-1}, y_{t-2}, \dots, y_{t-p}) \\ t = p+1, \dots, n$$

$$\text{AR: } \mu_t = \beta_0 + \beta^T x_t$$

$$\text{Loss: } \sum (y_t - \mu_t)^2$$

$$y_t = \mu_t + \varepsilon_t \\ \varepsilon_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

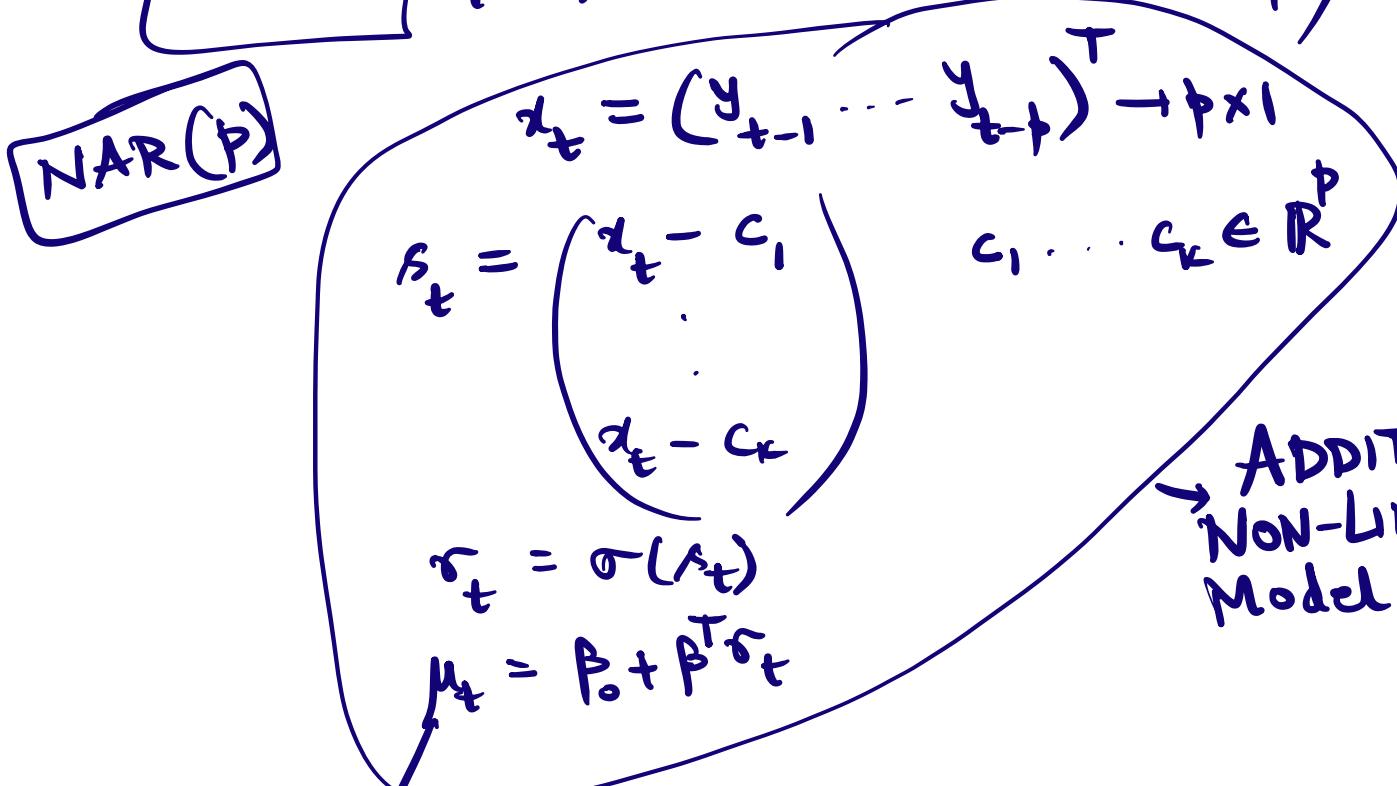
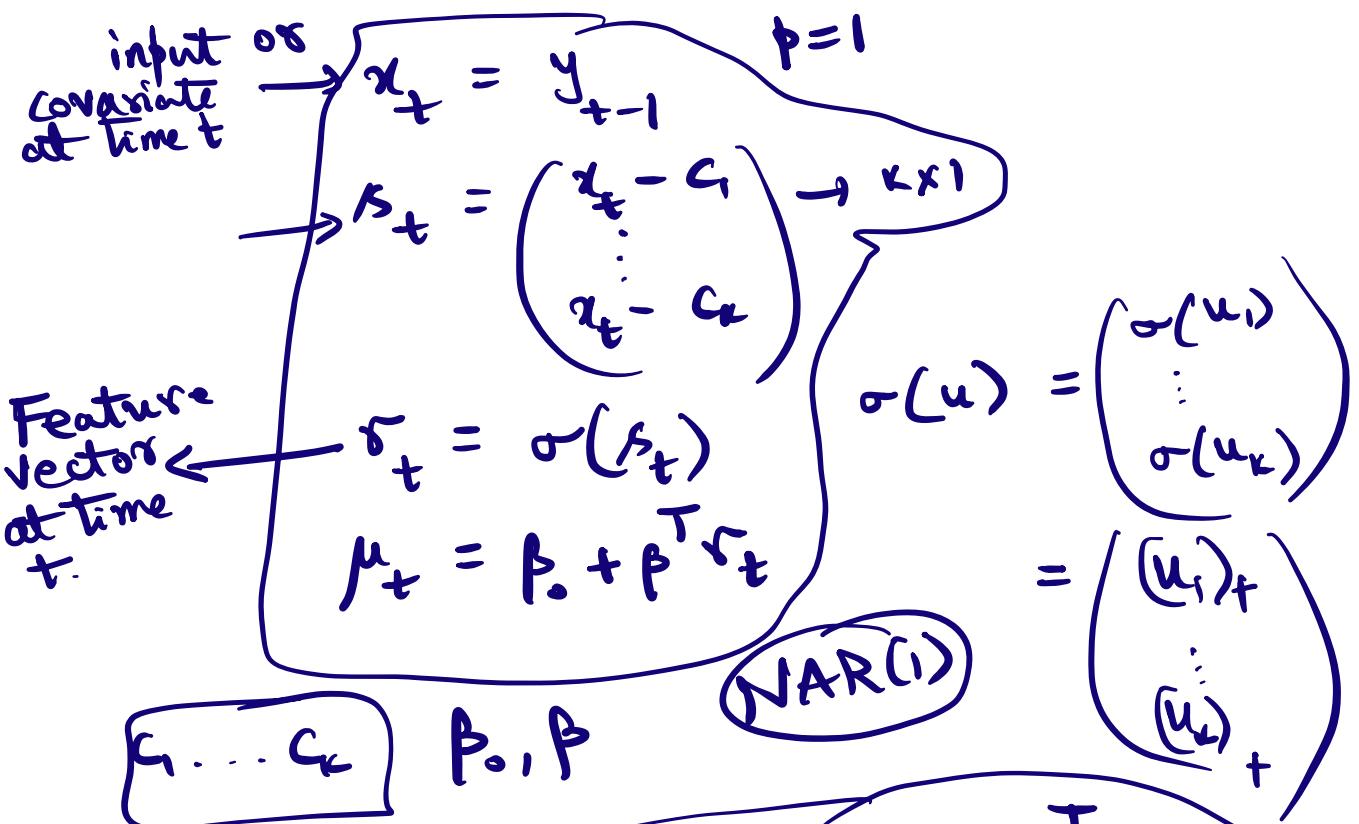
Qn: How to make AR nonlinear?

$$p=1 \quad x_t = y_{t-1} \\ \mu_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 (y_{t-1} - c_1)_+ + \dots + \beta_{k+1} (y_{t-1} - c_k)_+$$

$$\beta_0, \beta_1, \beta_2, \dots, \beta_{k+1}$$

$$y_{t-1} = y_{t-1} - c_0 + c_0 = (y_{t-1} - c_0)_+ + c_0$$

$$\mu_t = \beta_0 + \beta_1 (y_{t-1} - c_1)_+ + \dots + \beta_k (y_{t-1} - c_k)_+$$

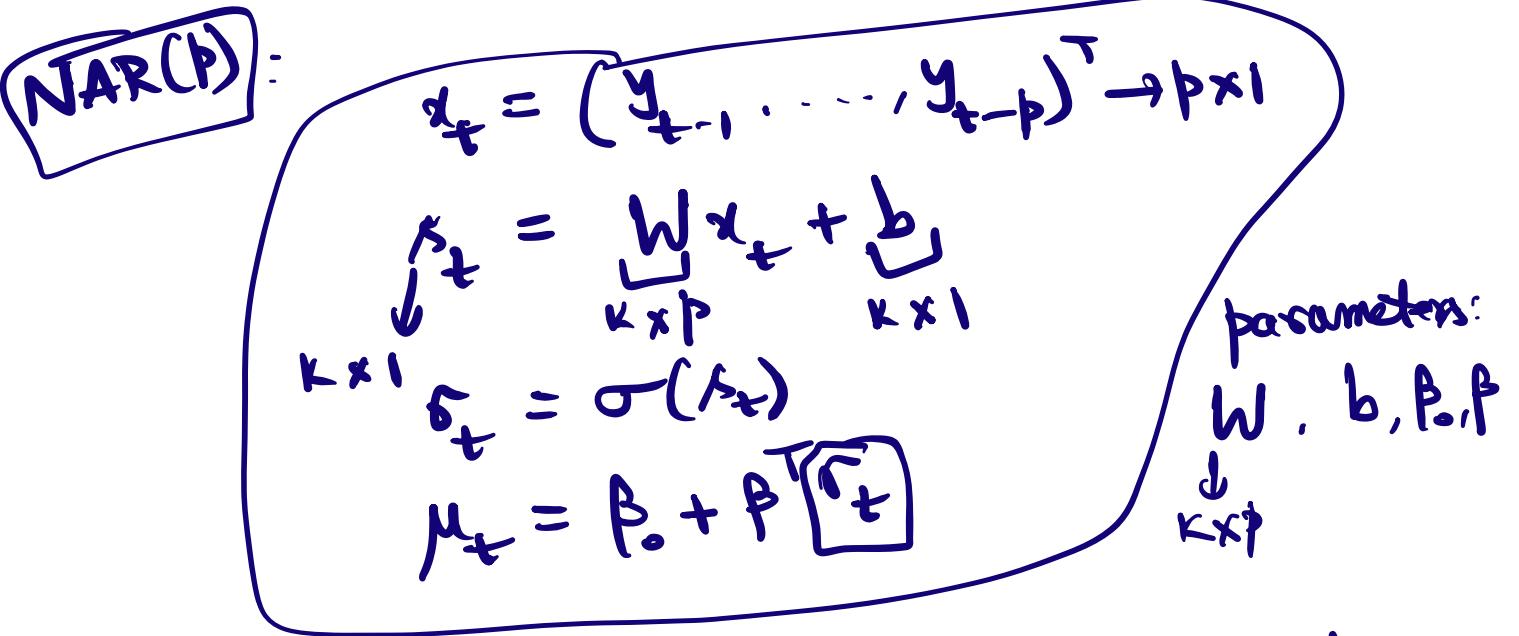


Check:

$$\mu_t = \left[\left(x_t^{(1)} - c_1 \right)_+ + \dots + \left(x_t^{(1)} - c_p \right)_+ \right] + \left[\left(x_t^{(2)} - c_1 \right)_+ + \dots + \left(x_t^{(2)} - c_p \right)_+ \right] + \left[\left(x_t^{(3)} - c_1 \right)_+ + \dots + \left(x_t^{(3)} - c_p \right)_+ \right]$$

Additive
model

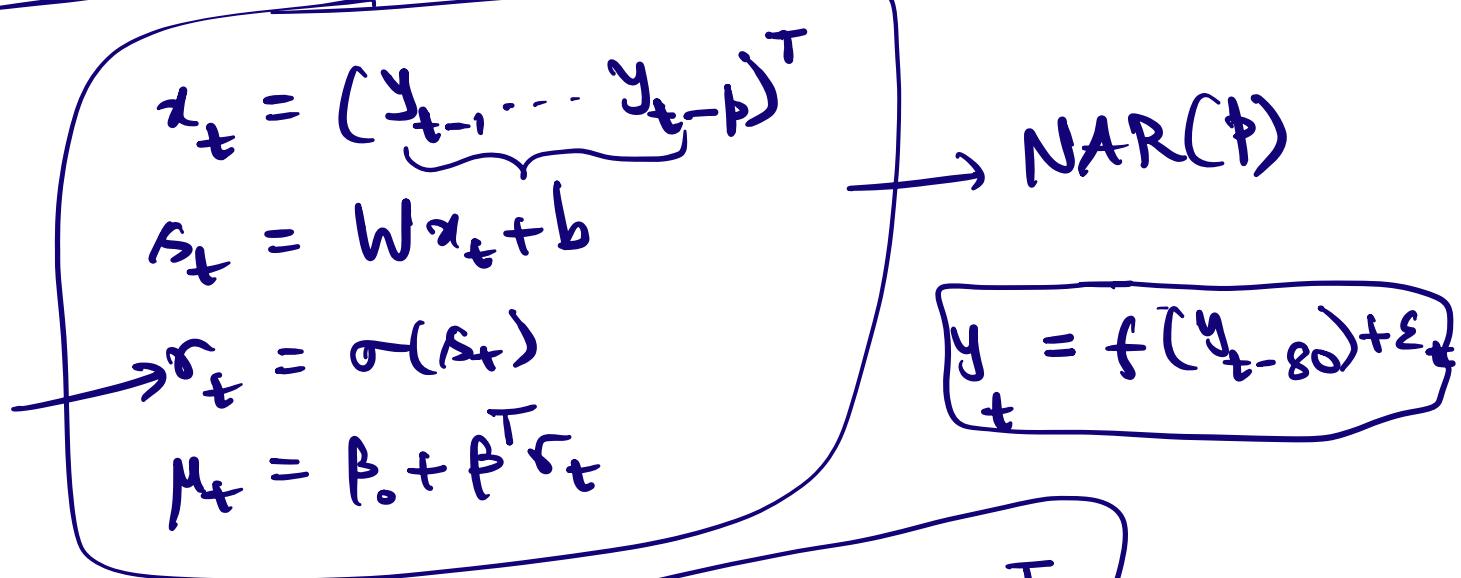
μ_t $y_{t-1} + y_t$



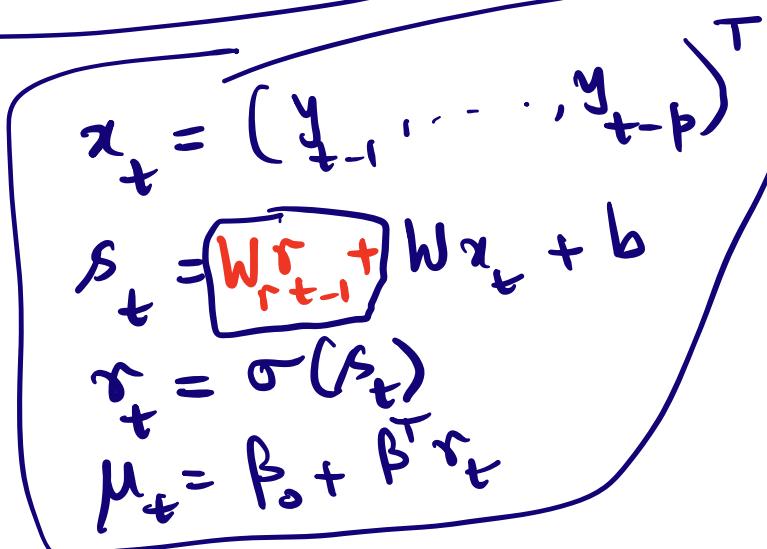
μ_t : depends on x_t because r_t depends on x_t

$$r_t = \sigma(W x_t + b)$$

Recurrent Neural Network (RNN)



RNN



RNN

$$x_t = (y_{t-1} \dots y_{t-p})^T$$

$$r_t = \sigma(W_r r_{t-1} + W_x x_t + b), r_0 = 0$$

$$\mu_t = \beta_0 + \beta^T r_t$$

(t=1)

$$r_1 = \sigma(W_r x_1 + b)$$

$$r_2 = \sigma(W_r r_1 + W_x x_2 + b)$$

$$= \sigma(W_r \sigma(W_r x_1 + b) + W_x x_2 + b)$$

→ depends on both x_2 & x_1

$$r_3 = \sigma(W_r r_2 + W_x x_3 + b)$$

$$= \sigma(W_r \sigma(W_r \sigma(W_r x_1 + b) + W_x x_2 + b) + W_x x_3 + b)$$

→ depends on x_3, x_2, x_1

More generally,

r_t depends on $x_t, x_{t-1}, x_{t-2}, \dots, x_1$

It can have stability issues.

$$\sigma(u) = \max(u, 0)$$

$$\tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} \in (-1, 1)$$

$$r_t = \tanh(W_r r_{t-1} + W_x x_t + b), \mu_t = \beta_0 + \beta^T r_t$$

We want W_f to not be more than 1.

(spectral radius of W_f should be < 1) largest modulus of any eigenvalue

But when W_f has spectral radius < 1 ,
 δ_t might have weak dependence on x_u when u is much smaller than t .

To verify this, calculate $\frac{\partial \delta_t}{\partial x_u}$ ^{$k \times 1$} JACOBIAN

$$\frac{\partial \delta_3}{\partial x_3} = \frac{\partial}{\partial x_3} \sigma \underbrace{(W_f \delta_2 + W x_3 + b)}_{s_3} = \sigma'(\beta_3) \frac{\partial \beta_3}{\partial x_3} = \sigma'(\beta_3) W$$

Chain rule
diagonal matrix $\begin{bmatrix} \sigma'(\beta_3(1)) \\ \vdots \\ \sigma'(\beta_3(k)) \end{bmatrix}$

$$\begin{aligned} \frac{\partial \delta_3}{\partial x_2} &= \frac{\partial}{\partial x_2} \sigma \underbrace{(W_f \delta_2 + W x_3 + b)}_{s_3} \\ &= \sigma'(\beta_3) \frac{\partial \beta_3}{\partial x_2} \\ &= \sigma'(\beta_3) W_f \frac{\partial \delta_2}{\partial x_2} = \sigma'(\beta_3) W_f \sigma'(\beta_2) W \end{aligned}$$

$$\frac{\partial r_3}{\partial x_1} = \sigma'(r_3) W_r \sigma'(r_2) W_r \sigma'(r_1) W$$

$$\frac{\partial r_{100}}{\partial x_{60}} = \sigma'(r_{100}) W_r \sigma'(r_{99}) W_r \dots \sigma'(r_{60}) W$$

$$\sigma(u) = \frac{e^{-u}}{e^u + e^{-u}} \Rightarrow \sigma'(u) = 1 - \sigma^2(u) \approx 0.1$$

very small

$\frac{\partial r_t}{\partial x_u}$ is probably very small when $u \ll t$

RNNs may not have LONG RANGE DEPENDENCE in practice.

$$r_t = \sigma(W_r r_{t-1} + W_x x_t + b), \quad r_0 = 0$$

Gated Recurrent Unit (GRU)

$$\tilde{r}_t = \tanh(W_r r_{t-1} + W_x x_t + b)$$

a) $r_t = \tilde{r}_t \rightarrow$ RNN (either explosion or lack of long range dependence)

b) $r_t = r_{t-1} \rightarrow$ will not use the current input x_t .

GRU:

$$\tilde{r}_t = \underbrace{z_t \odot r_{t-1}}_{K \times 1} + (1 - z_t) \tilde{r}_t, \quad t = 1, \dots, n$$

$$z_t = \sigma(W_{rz} \tilde{r}_{t-1} + W_z x_t + b_z)$$

sigmoid $\sigma(u) = \frac{e^u}{1+e^u} \in (0, 1)$

GRU:

$$\begin{aligned} \tilde{r}_t &= \tanh(W_r \tilde{r}_{t-1} + W_r x_t + b_r) \\ z_t &= \sigma(W_{rz} \tilde{r}_{t-1} + W_z x_t + b_z) \\ \tilde{r}_t &= z_t \odot r_{t-1} + (1 - z_t) \tilde{r}_t \\ \mu_t &= \beta_0 + \beta^T r_t \end{aligned}$$

z_t : UPDATE GATE

GRU:

$$\begin{aligned} \tilde{r}_t &= \tanh(W_r (\tilde{r}_{t-1} \odot g_t) + W_r x_t + b_r) \\ z_t &= \sigma(W_{rz} \tilde{r}_{t-1} + W_z x_t + b_z) \\ \tilde{r}_t &= z_t \odot r_{t-1} + (1 - z_t) \tilde{r}_t \\ g_t &= \sigma(W_{rg} \tilde{r}_{t-1} + W_g x_t + b_g) \\ \mu_t &= \beta_0 + \beta^T r_t \end{aligned}$$

UPDATE GATE

RESET GATE

$$r_t \rightarrow r_{t-1} \quad \alpha_t$$

$$\tilde{r}_t = \tanh(W_r r_{t-1} + W_x x_t + b)$$

$$r_t = z_t \odot r_{t-1} + (1 - z_t) \odot \tilde{r}_t$$

$$z_t = 0.9999$$

LSTM (Long Short Term Memory)

$$\tilde{r}_t = \tanh(W_r r_{t-1} + W_x x_t + b)$$

$$r_{t-1} \rightarrow r_t : \text{GRU}$$

$$r_{t-1} \rightarrow r_t : \text{RNN}$$

$$(r_{t-1}, \beta_{t-1}) \rightarrow (r_t, \beta_t)$$

$$s_t = \underbrace{\text{gate} \odot r_{t-1}}_{\text{forget}} + \underbrace{\text{another gate} \odot s_{t-1}}_{\text{update}}$$

$$r_t = \text{gate} \odot \sigma(s_t)$$