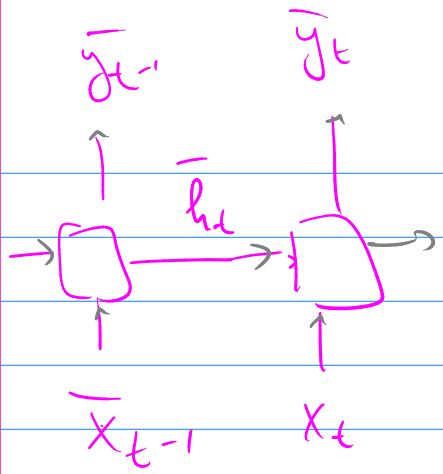
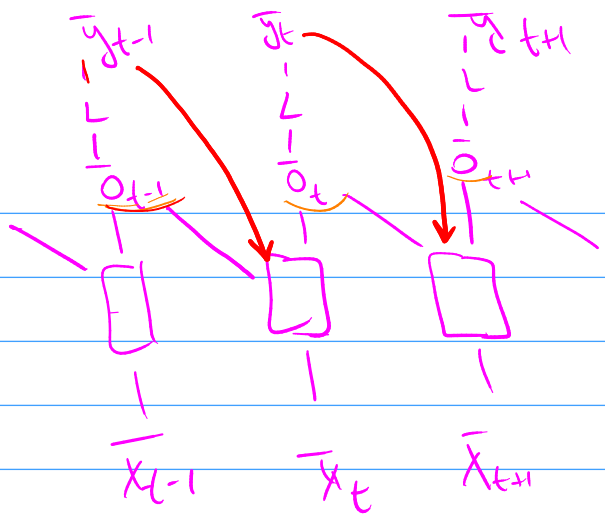
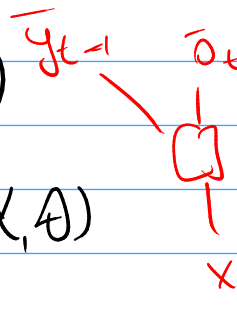




$$p(\bar{y}_1, \dots, \bar{y}_T \mid \bar{x}_1 \dots \bar{x}_T, \bar{\theta}) =$$

$$= p(\bar{y}_1 \mid \bar{x}_1 \dots \bar{x}_T, \bar{\theta})$$
$$p(\bar{y}_2 \mid \bar{y}_1, X, \bar{\theta})\, p(\bar{y}_3 \mid \bar{y}_1, \bar{y}_2, X, \theta)$$
$$\dots\; p(\bar{y}_T \mid \bar{y}_1 \dots \bar{y}_{T-1}, X, \theta)$$

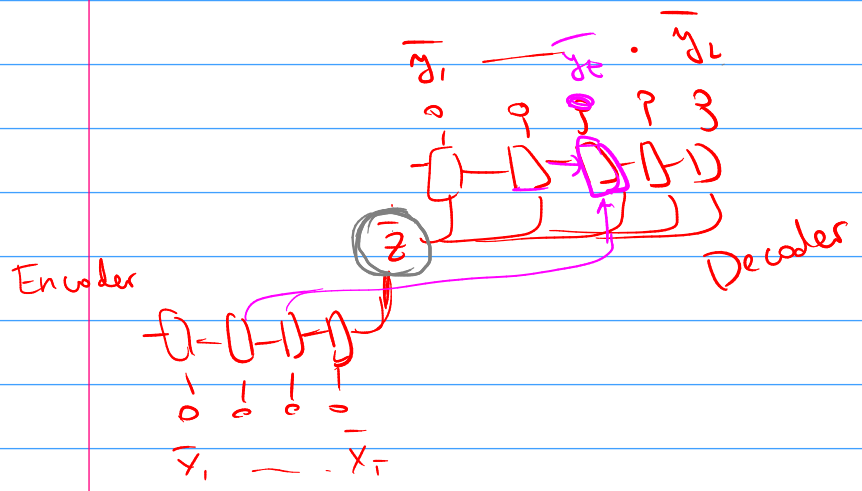
Jordan networks

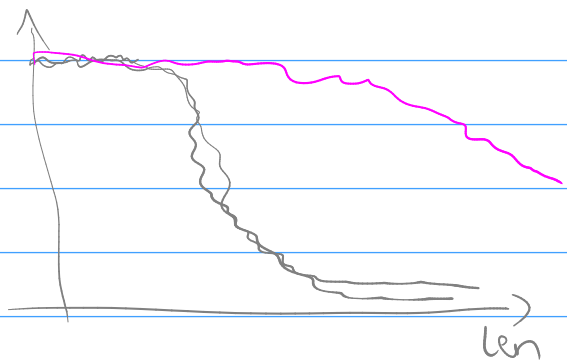


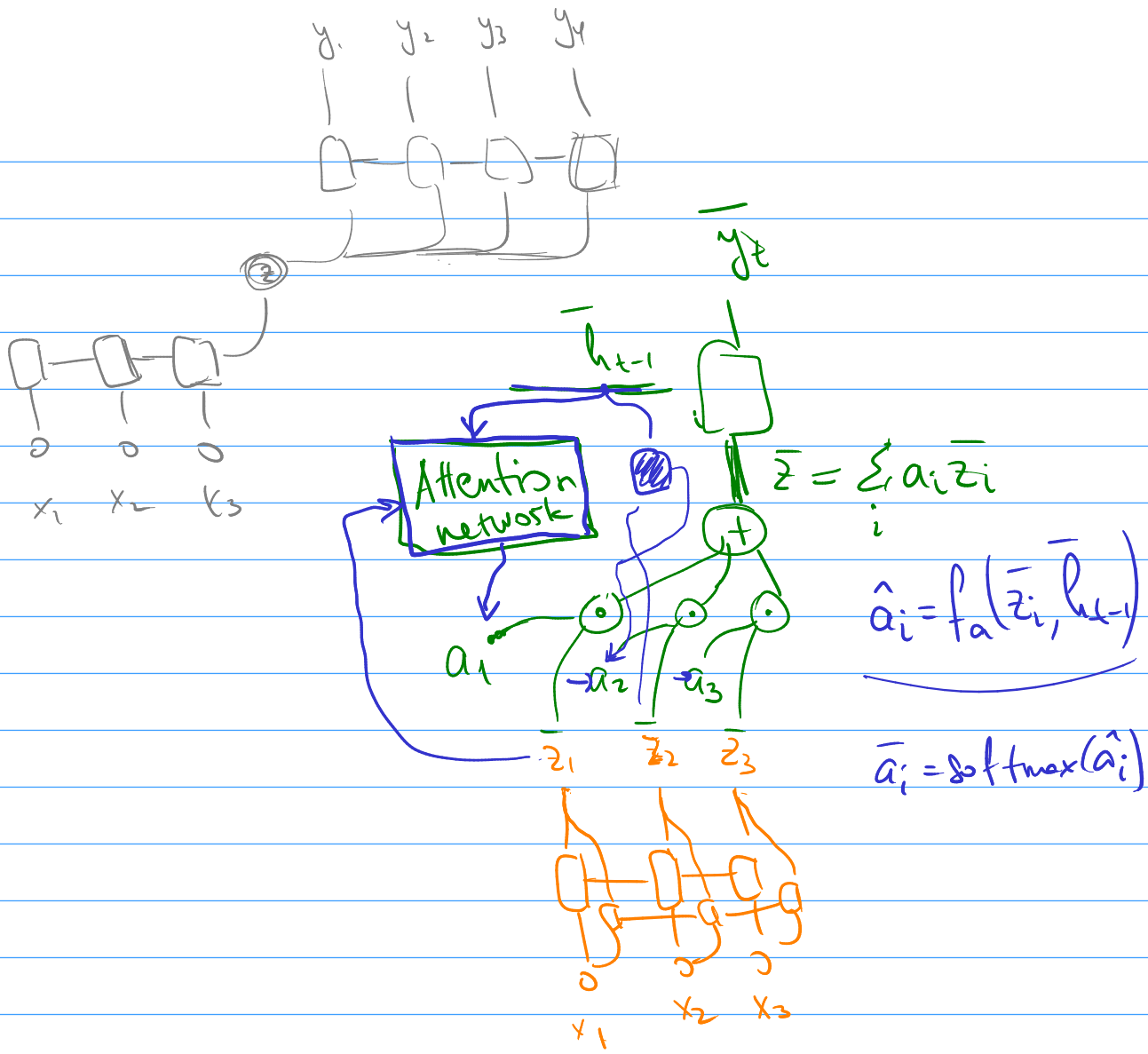
teacher forcing

$$p(\bar{y}_1 \dots \bar{y}_T \mid X, \theta) = p(\bar{y}_1 \mid \bar{x}_1, \bar{h}_0, \bar{\theta}) \cdot p(\bar{h}_1 \mid \bar{x}_1, \bar{h}_0, \bar{\theta}) \cdot$$
$$\cdot p(\bar{y}_2 \mid \bar{x}_2, \bar{h}_1, \bar{\theta})\, p(\bar{h}_2 \mid \bar{x}_2, \bar{h}_1, \bar{\theta}) \dots$$

$$p(\bar{y}_1 \dots \bar{y}_T \mid X, \theta) = p(\bar{y}_1 \mid \bar{x}_1, \bar{\theta})\, p(\bar{y}_2 \mid \bar{x}_1, \bar{x}_2, \bar{y}_1, \bar{\theta}) \dots$$






y_1 y_2 y_3 y_4
x_1 x_2 x_3
$\bar{y}_t$
$\bar{h}_{t-1}$
Attention network
$\bar{z} = \sum_i a_i \bar{z}_i$
$\hat{a}_i = f_a(\bar{z}_i, \bar{h}_{t-1})$
$\bar{a}_i = softmax(\hat{a}_i)$
a_1 a_2 a_3
$\bar{z}_1$ $\bar{z}_2$ $\bar{z}_3$
x_1 x_2 x_3