Ex 7

7.2) Transition feature template: $\langle y_t, y_{t-1} \rangle$

$$\Rightarrow \phi_1^{trans} = [[y_t = 1, y_{t-1} = 1]]$$

$$\phi_2^{trans} = [[y_t = 1, y_{t-1} = 2]]$$

$$\phi_3^{trans} = [[y_t = 2, y_{t-1} = 1]]$$

$$\phi_4^{trans} = [[y_t = 2, y_{t-1} = 2]]$$

Observation feature template: $\langle y_t, x_t \rangle$

$$\Rightarrow \phi_1^{obs} = [[y_t = 1, x_t = \text{"We"}]]$$

$$\phi_2^{obs} = [[y_t = 1, x_t = \text{"are"}]]$$

$$\phi_3^{obs} = [[y_t = 1, x_t = \text{"clever"}]]$$

$$\phi_4^{obs} = [[y_t = 2, x_t = \text{"We"}]]$$

$$\phi_5^{obs} = [[y_t = 2, x_t = \text{"are"}]]$$

$$\phi_6^{obs} = [[y_t = 2, x_t = \text{"clever"}]]$$

7.3) Use the result that for $a \geq 0$,

$$\max(a+b,\ a+c) = a + \max(b, c)$$

Let $a(y_{t-1}, y_t) = \sum_{k=1}^{d_{trans}} w_k^{trans} \phi_k^{trans}(y_t, y_{t-1})$

$$b(y_t, x, t) = \sum_{k=1}^{d_{obs}} w_k^{obs} \phi_k^{obs}(x, y_t)$$

Searching for : $\max_{y_1, \dots y_T} \left[\sum_{t=2}^{T} a(y_{t-1}, y_t) + \sum_{t=1}^{T} b(y_t, x, t) + \underbrace{\log Z(x)}_{\text{constant w.r.t 'y' -- can ignore for maximization.}} \right]$

$$\rightarrow \max_{y_1, \dots y_T} \left[\sum_{t=2} a(y_{t-1}, y_t) + \sum_{t=2} b(y_t, x, t) + \underbrace{b(y_1, x, 1)}_{= \delta_1(y_1)} \right]$$

$$= \max_{y_2, \dots y_T} \left[\sum_{t=3}^{T} a(y_{t-1}, y_t) + \sum_{t=3}^{T} b(y_t, x, t) + \underbrace{\max_{y_1} \left(\delta_1(y_1) + a(y_1, y_2) + b(y_2, x, 2) \right)}_{\delta_2(y_2) \text{ ... not a function of } y_1 \text{ anymore}} \right]$$

$\Psi_2(y_2)$: store argmax y_1 per y_2

$$= \max_{y_3 \cdots y_T} \left[\sum_{t=4}^{T} a(y_{t-1}, y_t) + \sum_{t=4}^{T} b(y_t, x, t) + \underbrace{\max_{y_2} \left(\delta_2(y_2) + a(y_2, y_3) + b(y_3, x, 3) \right)}_{\delta_3(y_3)} \right]$$

$\psi_3(y_3)$: store argmax y_2 per y_3

$= \cdots$

[Note that $\delta_i(\cdot)$ and $\psi_i(\cdot)$ are discrete functions with domain and range 'Σ' each, such that $\forall t : y_t \in \Sigma$]

Now, let $\Sigma = \{1, \cdots N\}$, w/o loss of generality.

then, for $i = 1, \cdots N$

$$\delta_1(i) = b(i, x, 1)$$

$$\psi_1(i) = -1$$ --- (not defined/unused)

for $t = 2, \cdots T$

for $j = 1, \cdots N$

$$\delta_t(j) = \max_i \left(\delta_{t-1}(i) + a(i,j) + b(j, x, t) \right)$$

$$\psi_t(j) = \underset{i}{\operatorname{argmax}} \left(\delta_{t-1}(i) + a(i,j) + \underbrace{b(j,x,t)}_{\text{constant w.r.t. } i,\ \text{can be ignored}} \right)$$

$$\hat{y}_T = \underset{i}{\operatorname{argmax}}\ \delta_T(i)$$

for $t = T-1, \dots 1$.

$$\hat{y}_t = \psi_{t+1}(\hat{y}_{t+1})$$