Computational Mechanics:

3. Prediction

ICIAD INTENSIVE - SPRING 2026

Outline:

1) Belief state

2) Next-token (probability) vector

3) Mixed-state presentation

4) GHMM prediction

Belief state

How should we predict tokens emitted from an HMM?

↳ By def^{n} the next-token distribution is a function of the current distribution over hidden states.

↳ Use this to make predictions

Bayes' rule tells us that the posterior prob. of occupying state s_j after observing x is:

$$P(s_j \mid x) = \frac{P(x, s_j)}{P(x)} = \frac{\sum_{s_0 \in S} P(s_0)\, P(x, s_j \mid s_0)}{P(x)}$$

s_j: hidden state; x: observed tokens

Similarly, posterior prob. of occupying s_j after observing the sequence $W = w_1 w_2 \cdots w_n$ is:

$$P(s_j | W) = \frac{\sum_{s^0, \ldots, s^{n-1}} P(s_0) P(w_1, s^1 | s^0) P(w_2, s^2 | s^1) \cdots P(w_n, s_j | s^{n-1})}{P(W)}$$

Expressing this object in terms of HMM data:

$$P(s_j | W) = \left(\frac{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)}}{\underbrace{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)} \mathbf{1}}_{\text{row-vector}}} \right)_j$$

where we order the hidden states $S = \{s_0, s_1, \ldots\}$.

This motivates the def^{n} of the <u>belief state</u>:

$$\eta^{(w)} := \frac{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)}}{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)} \mathbf{1}}, \qquad w = w_1 w_2 \cdots w_n \in \mathcal{X}^*$$

whose j^{th}-element corresponds to the prob. that the hidden state is S_j after obs. w.

Suppose we observe a further token $x \in \mathcal{X}$, the predictive vector update rule is simple:

$$\eta^{(w)} \longmapsto \eta^{(wx)} = \frac{\eta^{(w)} T^{(x)}}{\eta^{(w)} T^{(x)} \mathbf{1}}$$

Next-token (probability) vector

For predicting next-token emissions, we actually care about:

$$P(x_j \mid w) = \frac{P(w x_j)}{P(w)} = \frac{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)} T^{(x_j)} \mathbf{1}}{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)} \mathbf{1}}$$

x_j: next-token; w: observed tokens

This motivates the def^{n} of the next-token vector:

$$p^{(w)} := \left(\frac{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)} T^{(x_j)} \mathbf{1}}{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)} \mathbf{1}} \right)_{j=0,1,\ldots}$$

where we order the tokens $\mathcal{X} = \{x_0, x_1, \ldots\}$.

The j^{th} element of $p^{(w)}$ corresponds to the prob. that the next-token is x_j after obs. w.

The belief state appears directly in the def^{n} of the next-token vector:

$$p_j^{(w)} := \frac{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \dots T^{(w_n)} T^{(x_j)} \mathbf{1}}{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \dots T^{(w_n)} \mathbf{1}} = \eta^{(w)} T^{(x_j)} \mathbf{1}$$

It follows that there exists a matrix that maps $\eta^{(w)} \mapsto p^{(w)}$ i.e.:

$$p^{(w)} = \eta^{(w)} A$$

where the matrix elements of A are:

$$A_{ij} := \sum_{k=0}^{|S|-1} T_{ik}^{(x_j)}, \qquad \begin{matrix} i = 0, 1, \dots |S|-1 \\ j = 0, 1, \dots |X|-1 \end{matrix}.$$

$$\mathcal{M}^{(w)} \xrightarrow{A} \mathcal{P}^{(w)}$$

Hidden dynamic

Observations

If $\ker(A) = \varnothing$:

If $\ker(A) \neq \varnothing$:

Belief states are also useful for all future preds.

Let $w^{(p)} = w_1^{(p)} w_2^{(p)} \cdots w_m^{(p)}$ & $w^{(f)} = w_1^{(f)} w_2^{(f)} \cdots w_n^{(f)}$:

$$P(w^{(f)} | w^{(p)}) = \frac{\eta^{(\phi)} T^{(w_1^{(p)})} \cdots T^{(w_m^{(p)})} T^{(w_1^{(f)})} \cdots T^{(w_n^{(f)})} \mathbf{1}}{\eta^{(\phi)} T^{(w_1^{(p)})} \cdots T^{(w_m^{(p)})} \mathbf{1}}$$

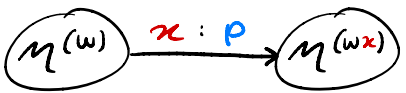
$$= \eta^{(w^{(p)})} T^{(w_1^{(f)})} \cdots T^{(w_n^{(f)})} \mathbf{1}$$

As in the next-token case there exists a multi-linear map from $\eta^{(w^{(p)})}$ to all future predictions.

Mixed-state presentation

For any HMM, the deterministic update rule $\eta^{(w)} \mapsto \eta^{(wx)}$ defines a meta-dynamic over beliefs called the mixed-state presentation (MSP).

The MSP is itself a (unifilar) HMM:



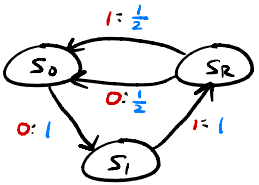
$p = P(x|w)$

whose hidden states are the belief states.

[For unifilar HMMs, an ε-machine is the MSP of the corresponding minimal HMM.]

Example 1: (zero-one-random $p = \frac{1}{2}$)

Recall:



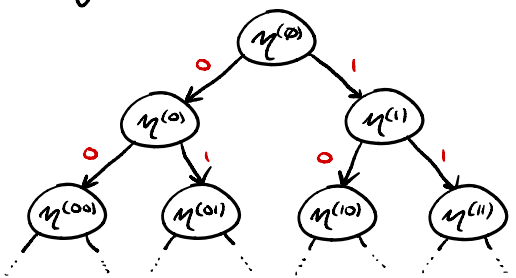
$$T^{(0)} = \begin{array}{c} \begin{matrix} S_0 & S_1 & S_R \end{matrix} \\ \left[\begin{matrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{matrix}\right] \end{array} \begin{matrix} S_0 \\ S_1 \\ S_R \end{matrix}$$

$$T^{(1)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 \end{bmatrix}$$

$$\eta^{(w)} = \frac{\eta^{(\phi)} T^{(w)}}{\eta^{(\phi)} T^{(w)} \mathbf{1}}$$

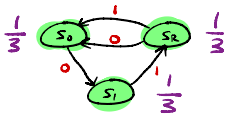
$$T^{(w)} := T^{(w_1)} \cdots T^{(w_n)}$$

Deriving the MSP compute:



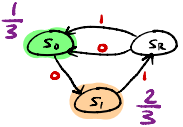
(identify nodes where $\eta^{(w)} = \eta^{(w')}$)

Fix: $\mu^{(\phi)} = [\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}]$

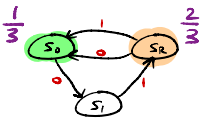


and compute:

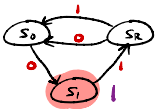
$\mu^{(0)} = [\frac{1}{3} \quad \frac{2}{3} \quad 0]$



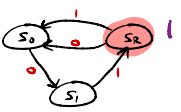
$\mu^{(1)} = [\frac{1}{3} \quad 0 \quad \frac{2}{3}]$



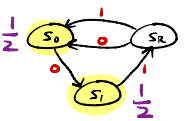
$\mu^{(00)} = [0 \quad 1 \quad 0]$



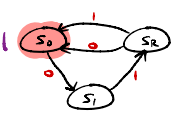
$\mu^{(01)} = [0 \quad 0 \quad 1]$



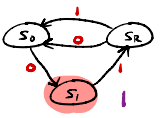
$\mu^{(10)} = [\frac{1}{2} \quad \frac{1}{2} \quad 0]$



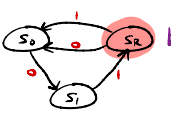
$\mu^{(11)} = [1 \quad 0 \quad 0]$



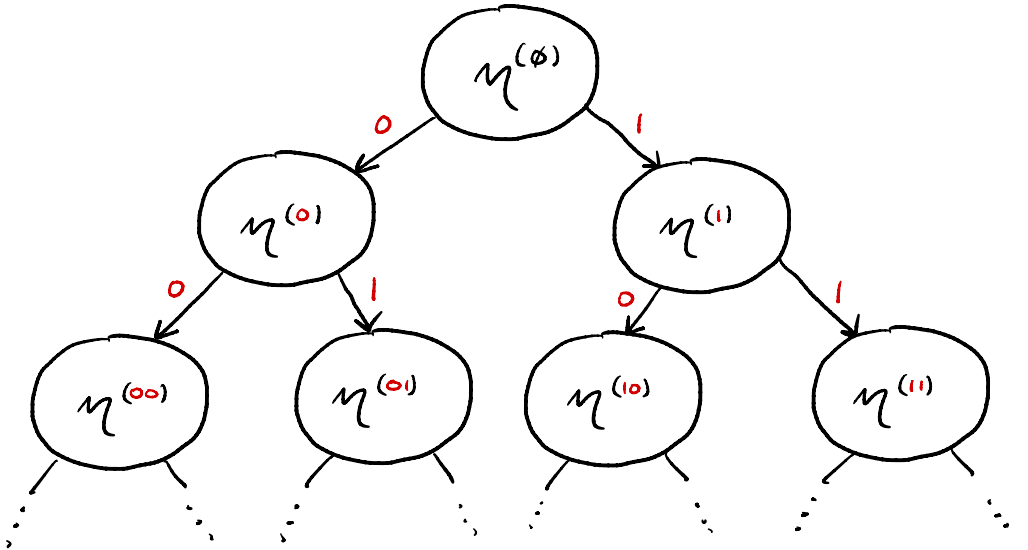
$\mu^{(100)} = [0 \quad 1 \quad 0]$

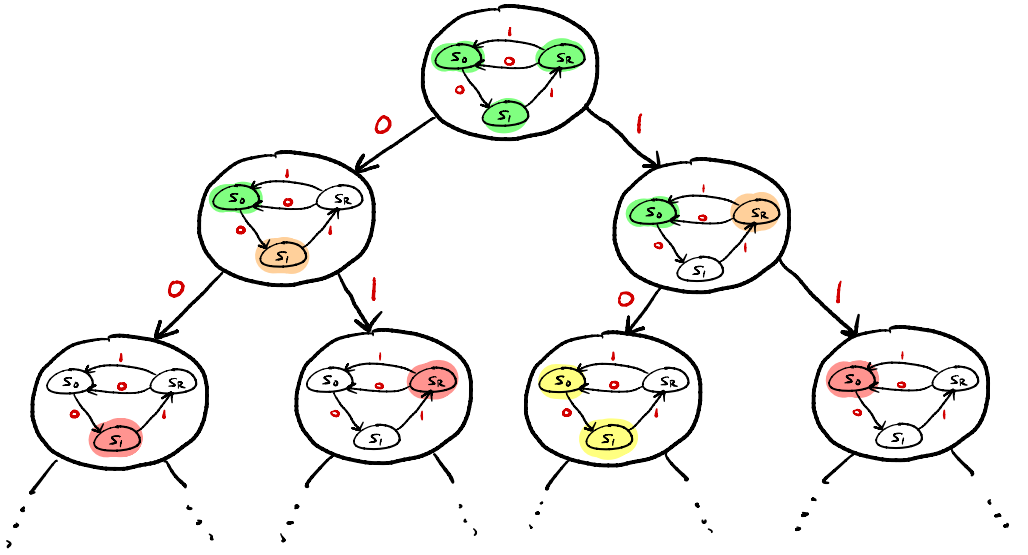


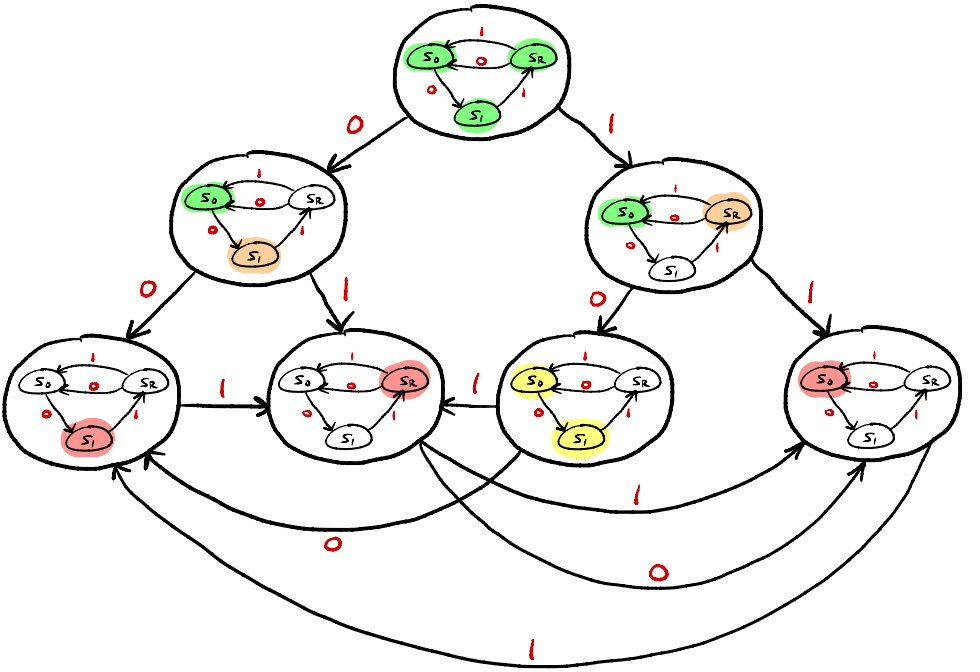
$\mu^{(101)} = [0 \quad 0 \quad 1]$

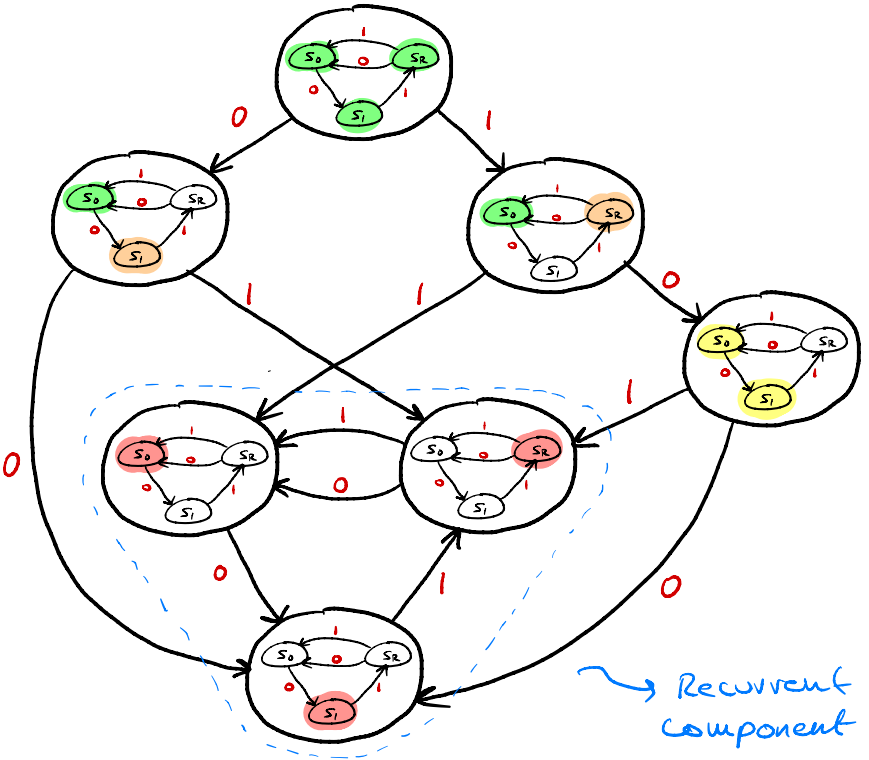


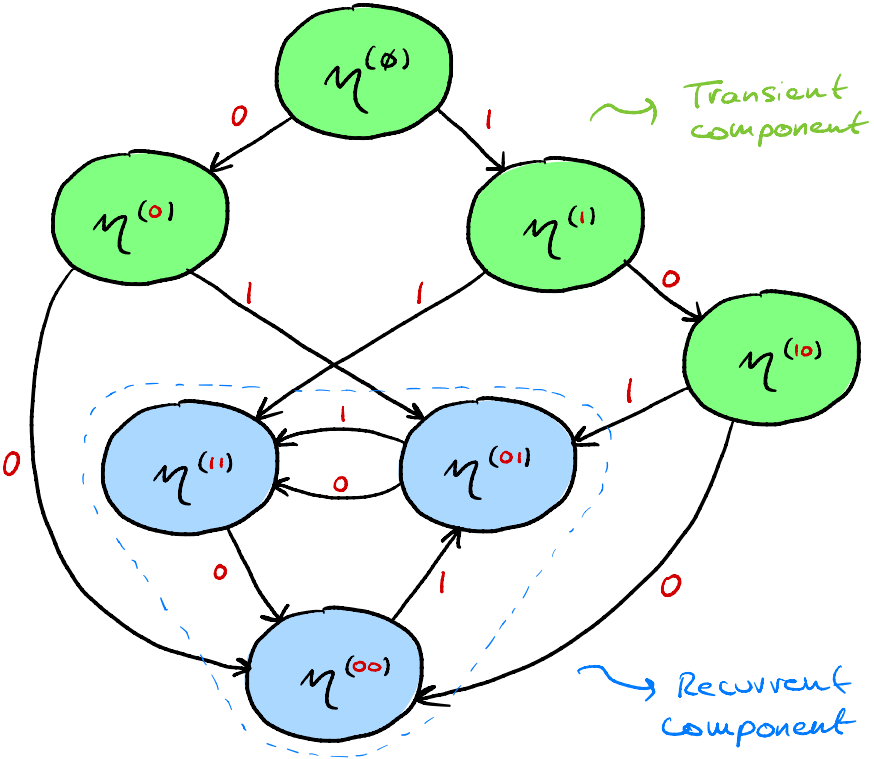
Inference process converges!


$\eta^{(\phi)}$
0
1
$\eta^{(0)}$
$\eta^{(1)}$
0
1
0
1
$\eta^{(00)}$
$\eta^{(01)}$
$\eta^{(10)}$
$\eta^{(11)}$


0
1
0
1
0
1
S_0
S_R
S_1


S_0
S_R
S_1
0
1

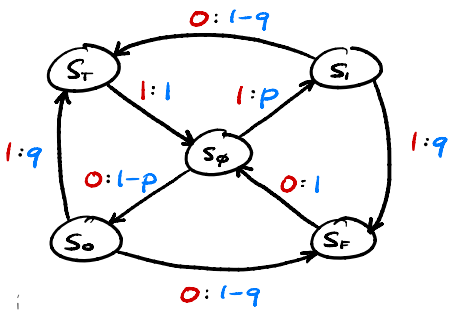

s_0
s_R
s_1
0
1
Recurrent component


$\eta^{(\phi)}$
$\eta^{(0)}$
$\eta^{(1)}$
$\eta^{(10)}$
$\eta^{(11)}$
$\eta^{(01)}$
$\eta^{(00)}$
0
1
Transient component
Recurrent component

Example 2: (random-random-XOR)

$$S = \{S_\phi, S_0, S_1, S_F, S_T\}$$

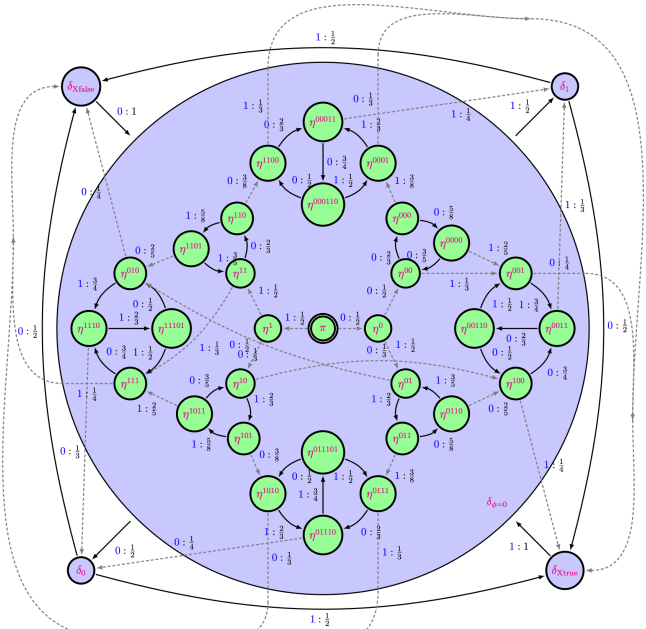
$$\chi = \{0, 1\}$$



$$T^{(0)} = \begin{array}{c} \\ S_\phi \\ S_0 \\ S_1 \\ S_F \\ S_T \end{array} \begin{array}{c} \begin{array}{ccccc} S_\phi & S_0 & S_1 & S_F & S_T \end{array} \\ \begin{bmatrix} 0 & 1-p & 0 & 0 & 0 \\ 0 & 0 & 0 & 1-q & 0 \\ 0 & 0 & 0 & 0 & 1-q \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

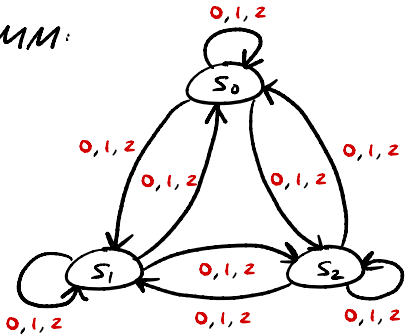
$$T^{(1)} = \begin{bmatrix} 0 & 0 & p & 0 & 0 \\ 0 & 0 & 0 & 0 & q \\ 0 & 0 & 0 & q & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The corresponding MSP:



Example 3: (Mess 3)

HMM:

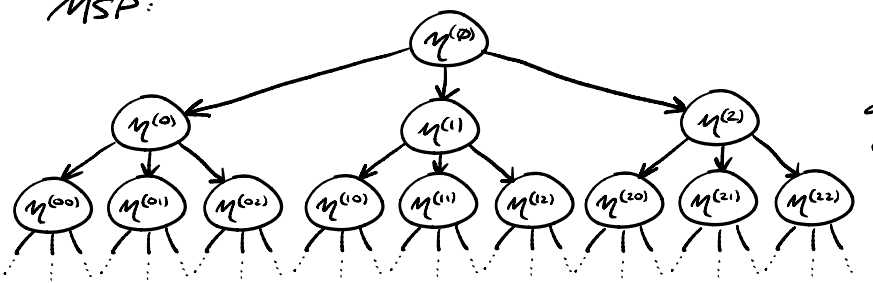


$$T^{(0)} = \begin{array}{c} \begin{matrix} S_0 & S_1 & S_2 \end{matrix} \\ \left[\begin{matrix} \alpha_y & \beta_x & \beta_x \\ \alpha_x & \beta_y & \beta_x \\ \alpha_x & \beta_x & \beta_y \end{matrix}\right] \end{array} \begin{matrix} S_0 \\ S_1 \\ S_2 \end{matrix}$$

$$T^{(1)} = \begin{bmatrix} \beta_y & \alpha_x & \beta_x \\ \beta_x & \alpha_y & \beta_x \\ \beta_x & \alpha_x & \beta_y \end{bmatrix}$$

$$T^{(2)} = \begin{bmatrix} \beta_y & \beta_x & \alpha_x \\ \beta_x & \beta_y & \alpha_x \\ \beta_x & \beta_x & \alpha_y \end{bmatrix}$$

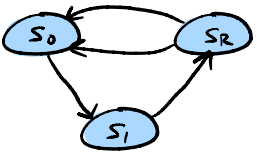
MSP:



Inference does not converge!

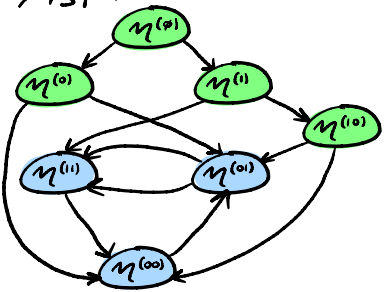
Summary:

HMM:



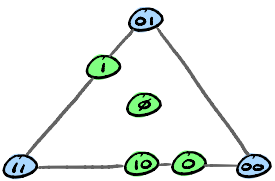
inference

MSP:



geometry

Belief-state geometry:



linear map

Next-token geometry:

GHMM prediction:

Throughout we have considered prediction of HMM data. Fortunately, all of the relevant objects have straightforward analogues in the generalised setting.

Recall that a GHMM consists of:

$(T^{(x)})_{x \in X}$ $\quad \eta^{(\phi)}$ – initial state $\quad \varepsilon$ – final state

Belief state (aka predictive vector):

$$\eta^{(w)} := \frac{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)}}{\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)} \varepsilon}, \qquad w = w_1 w_2 \cdots w_n \in X^*$$

Next-token vector:

$$p^{(w)} = \left(\eta^{(w)} T^{(x_j)} \varrho\right)_{j=0,1,\dots}$$

Linear map $\eta^{(w)} \xrightarrow{A} p^{(w)}$:

$$A_{ij} := \sum_{k=0}^{d-1} T_{ik}^{(x_j)} \varrho_k \qquad \begin{matrix} i = 0, 1, \dots, d-1 \\ j = 0, 1, \dots, |\mathcal{X}|-1 \end{matrix}.$$

A GHMM does not have 'underlying' hidden states so $\eta^{(w)}$ is not stochastic and thus cannot be viewed as a posterior distribution.

↪ Despite this the belief state remains similarly instrumental at making token predictions.

Questions?