Computational Mechanics:

2. Hidden Markov Models

ICIAD INTENSIVE - SPRING 2026

Outline:

1) What is a hidden Markov Model (HMM)?

2) HMM minimality & uniqueness

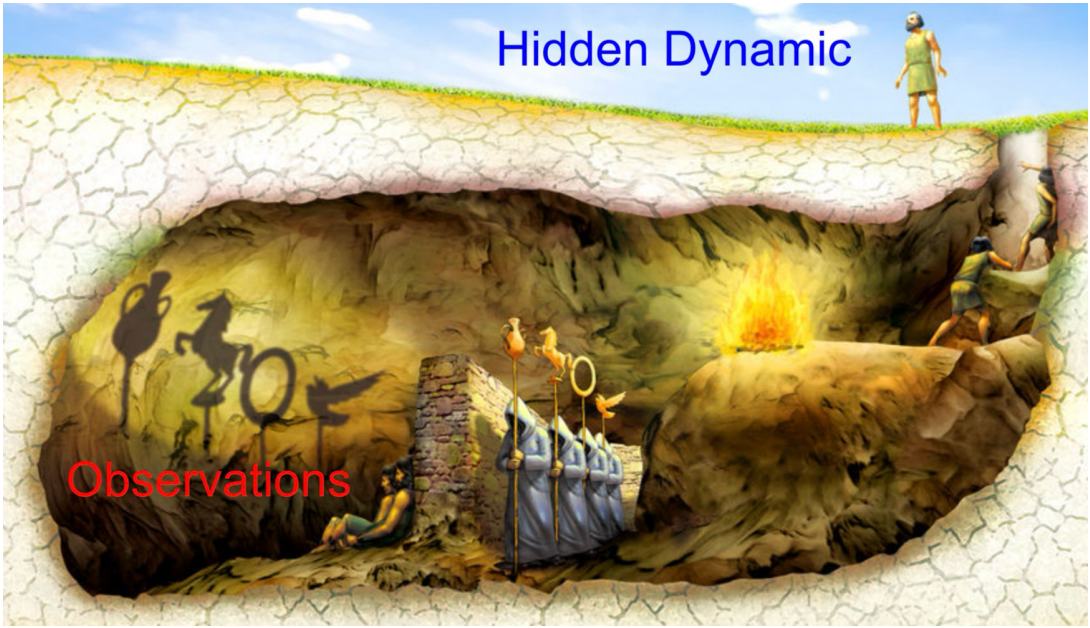
3) Generalised hidden Markov Model (GHMM)

What is a hidden Markov Model?

Intuitively, a hidden Markov Model (HMM) is a stochastic process that consists of two components:

* A 'hidden' unobserved component that evolves according to a Markovian dynamic — the next hidden state only depends on the current hidden state.
* An observed component that depends only on the current hidden state in a possibly lossy way.

What is a hidden Markov Model?



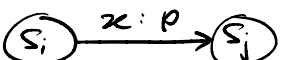
What is a hidden Markov Model?

A (Mealy-type) HMM consists of:

* A set S "think" ~~~ collection of hidden states
* A set $\mathcal{X}$ "think" ~~~ vocabulary of observations
* A collection of transition matrices $(T^{(x)})_{x \in \mathcal{X}}$ "think" ~~~ dynamic that determines emissions & hidden state transitions

where $T^{(x)}_{ij} := P(X = x, S_j = s_j \mid S_i = s_i) = P(x, s_j \mid s_i)$

It is convenient to use the following graphical notation to visualise process transitions:



, $\quad p = P(x, s_j \mid s_i)$

Fix an initial distribution over hidden states

$$\eta^{(\phi)}_i := (P(s_i))_i, \quad i = 1, \dots, |S|$$

The probability of observing the sequence of emissions $w := w_1 w_2 \cdots w_n \in \mathcal{X}^n$ factorises as:

$$P(w) = \sum_{s_0, \dots, s_n} P(s_0) P(w_1, s_1 | s_0) P(w_2, s_2 | s_1) \cdots P(w_n, s_n | s_{n-1})$$

Conveniently, the above expression can be stated in terms of HMM transition matrices as:

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

where $\eta^{(\phi)}$ – row vector w.r.t ordered basis of S & $\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

Example 1: (Biased coin)

$S = \{s_0\}$

$\mathcal{X} = \{0, 1\}$

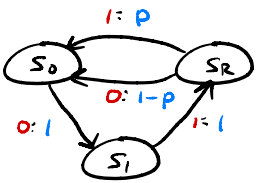
$$T^{(0)} = 1 - p$$

$$T^{(1)} = p$$

Example 2: (zero-one-random)

$S = \{s_0, s_1, s_R\}$

$\mathcal{X} = \{0, 1\}$



$$T^{(0)} = \begin{array}{c} \begin{array}{ccc} s_0 & s_1 & s_R \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1-p & 0 & 0 \end{bmatrix} \end{array} \begin{array}{l} s_0 \\ s_1 \\ s_R \end{array}$$

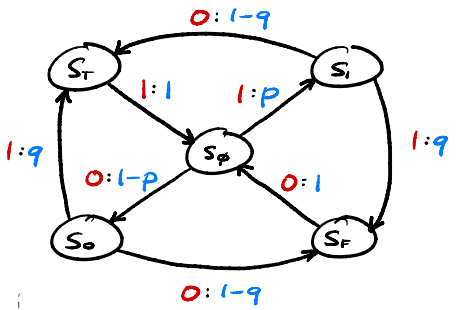
$$T^{(1)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ p & 0 & 0 \end{bmatrix}$$

$\ldots 0\,1\,R\,0\,1 \ldots \qquad R \in \{0, 1\}$

Example 3: (random-random-XOR)

$S = \{s_\phi, s_0, s_1, s_F, s_T\}$

$\mathcal{X} = \{0, 1\}$



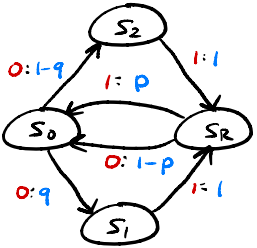
$$T^{(0)} = \begin{array}{c|ccccc} & s_\phi & s_0 & s_1 & s_F & s_T \\ \hline s_\phi & 0 & 1-p & 0 & 0 & 0 \\ s_0 & 0 & 0 & 0 & 1-q & 0 \\ s_1 & 0 & 0 & 0 & 0 & 1-q \\ s_F & 1 & 0 & 0 & 0 & 0 \\ s_T & 0 & 0 & 0 & 0 & 0 \end{array}$$

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

Example 4:

$$S = \{s_0, s_1, s_2, s_R\}$$

$$\mathcal{X} = \{0, 1\}$$



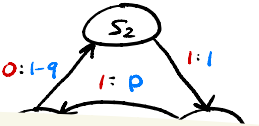
$$\tilde{T}^{(0)} = \begin{array}{c} \begin{matrix} s_0 & s_1 & s_2 & s_R \end{matrix} \\ \left[\begin{array}{cccc} 0 & q & 1-q & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1-p & 0 & 0 & 0 \end{array}\right] \end{array} \begin{matrix} s_0 \\ s_1 \\ s_2 \\ s_R \end{matrix}$$

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

Example 4:

$S = \{s_0, s_1, s_2\}$

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



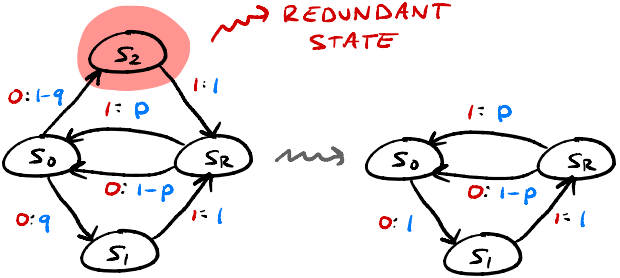
Same process as zero-one-random!

$$\tilde{T}^{(0)} = \begin{bmatrix} 0 & & & & \\ 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & \\ 1-p & 0 & 0 & 0 & \end{bmatrix} \begin{matrix} s_0 \\ \\ s_2 \\ s_2 \end{matrix} \begin{bmatrix} & & 0 & 0 \\ & & 0 & 1 \\ & & 0 & 1 \\ & & p & 0 \end{bmatrix}$$

Example 4:

$$S = \{s_0, s_1, s_2, s_R\}$$

$$\mathcal{X} = \{0, 1\}$$



$$\tilde{T}^{(0)} = \begin{array}{c} \begin{matrix} s_0 & s_1 & s_2 & s_R \end{matrix} \\ \begin{bmatrix} 0 & q & 1-q & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1-p & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} s_0 \\ s_1 \\ s_2 \\ s_R \end{matrix}$$

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

Let $(T^{(x)})_{x \in \mathcal{X}}$ – ZIR

$(\tilde{T}^{(x)})_{x \in \mathcal{X}}$ – ZIR with redundancy

There exist matrices U & V such that:

$$\tilde{T}^{(x)} = V T^{(x)} U \quad \& \quad Id_{3\times 3} = UV$$

$\eta^{(\phi)} U$ – stochastic & VT – all ones

for all $x \in \{0,1\}$. It follows that

$$\eta^{(\phi)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)} \mathbb{1}$$

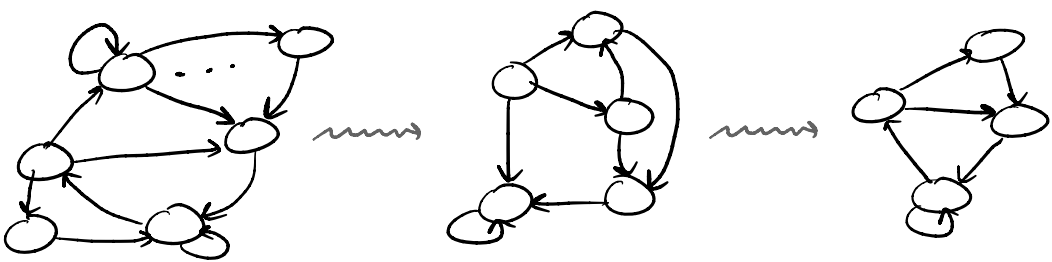
$$= \eta^{(\phi)} U V T^{(w_1)} U V T^{(w_2)} U \cdots V T^{(w_n)} U V \mathbb{1}$$

$$= \tilde{\eta}^{(\phi)} \tilde{T}^{(w_1)} \tilde{T}^{(w_2)} \cdots \tilde{T}^{(w_n)} \tilde{\mathbb{1}}$$

HMM minimality & uniqueness

A given HMM realisation of a stochastic process may be needlessly large i.e. have more hidden states than necessary.

We say that a HMM realisation of a stochastic process is <u>minimal</u> if there exists no other HMM realisations with fewer hidden states.



Claim: For a stochastic process that admits an HMM realisation, the minimal HMM is not unique.

(I.e. there does not exist similarity transforms relating the minimal HMMs.)

This is quite dissatisfying!

Suppose we trained a transformer on a process whose minimal HMM is not unique; what realisation would it learn the details of?

Can we identify a simple generalisation of HMMs such that uniqueness is guaranteed for the more general class?

Consider the output distribution of an HMM with $(T^{(x)})_{x\in\mathcal{X}}$ & initial hidden state dist. $\eta^{(\phi)}$

$$P(w) = \eta^{(\phi)} T^{(w_1)} T^{(w_2)} \dots T^{(w_n)} \mathbf{1}$$

$$= \eta^{(\phi)} V^{-1} V T^{(w_1)} V^{-1} V T^{(w_2)} V^{-1} \dots V T^{(w_n)} V^{-1} V \mathbf{1}$$

$$= \check{\eta}^{(\phi)} \check{T}^{(w_1)} \check{T}^{(w_2)} \dots \check{T}^{(w_n)} \check{\mathbf{1}}, \qquad V \in GL_{|S|}$$

where $\check{T}^{(x)} = V T^{(x)} V^{-1}$ & $\check{\eta}^{(\phi)} = \eta^{(\phi)} V^{-1}$ & $\check{\mathbf{1}} = V\mathbf{1}$

An identical process is produced by the transition matrices $(\check{T}^{(x)})_{x\in\mathcal{X}}$ and the vectors $\check{\eta}^{(\phi)}$ & $\check{\mathbf{1}}$.

$\Longrightarrow$ Why constrain the transition matrix to have probabilities as matrix elements?

Generalised hidden Markov Model

A generalised hidden Markov Model (GHMM) consists of:

* A set $\mathcal{X}$
* A collection of transition matrices $(T^{(x)})_{x \in \mathcal{X}}$ ⇝ not necessarily sub-Stochastic!
* An initial row vector $\eta^{(\phi)}$ ⇝ not necessarily Stochastic!

such that:

1. There exists a (column) vector φ satisfying:
   a) $T\varphi = \varphi$ where $T := \sum_{x \in \mathcal{X}} T^{(x)}$
   b) $\eta^{(\phi)} \varphi = 1$
2. For all $w \in \mathcal{X}^*$, $\eta^{(\phi)} T^{(w_1)} \cdots T^{(w_n)} \varphi \geq 0$.

The thrust of this definition is to afford additional freedom to $(T^{(x)})_{x \in \mathcal{X}}$ and $\eta^{(\emptyset)}$ while maintaining the same expression assigning probabilities to emission sequences:

$$P(w) = \eta^{(\emptyset)} T^{(w_1)} T^{(w_2)} \cdots T^{(w_n)} \varrho$$

(This fact will be explored in the exercises.)

A consequence of this freedom is that we lose connection to explicit hidden states!

Example: (Bloch Walk)

Fix $\gamma = (2\sqrt{\alpha^2+\beta^2})^{-1}$, where $\alpha > 0$ & $\beta \in \mathbb{R}$

$$T^{(0)} = \begin{bmatrix} \frac{1}{4} & 0 & 2\alpha\beta\gamma^2 \\ 0 & (\alpha^2-\beta^2)\gamma^2 & 0 \\ 2\alpha\beta\gamma^2 & 0 & \frac{1}{4} \end{bmatrix} \qquad T^{(1)} = \begin{bmatrix} \frac{1}{4} & 0 & -2\alpha\beta\gamma^2 \\ 0 & (\alpha^2-\beta^2)\gamma^2 & 0 \\ -2\alpha\beta\gamma^2 & 0 & \frac{1}{4} \end{bmatrix}$$

$$T^{(2)} = \begin{bmatrix} \frac{1}{4} & 2\alpha\beta\gamma^2 & 0 \\ 2\alpha\beta\gamma^2 & \frac{1}{4} & 0 \\ 0 & 0 & (\alpha^2-\beta^2)\gamma^2 \end{bmatrix} \qquad T^{(3)} = \begin{bmatrix} \frac{1}{4} & -2\alpha\beta\gamma^2 & 0 \\ -2\alpha\beta\gamma^2 & \frac{1}{4} & 0 \\ 0 & 0 & (\alpha^2-\beta^2)\gamma^2 \end{bmatrix}$$

$$\eta^{(\phi)} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \qquad \varphi = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Returning to our initial motivation...

Claim (Upper '97): For a stochastic process that admits a GHMM realisation, the minimal GHMM is unique (up to similarity transformation).

=> Transformers trained on data emitted from a GHMM should learn the structure corresponding to the minimal GHMM

We will discuss this 'structure' in the next block!

Questions?