ILIAD Intensive

Computational Mechanics

3. Representing the past

Xavier Poncini (Simplex)

Autumn 2026

Outline

1. Hidden Markov models
2. Belief states and next-token prediction
3. Sufficient statistics
4. Generalised HMMs and minimal-dimensional realisations

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**.
- An **observed** component that provides a possibly lossy view of the hidden dynamic.


Hidden Dynamic
Observations

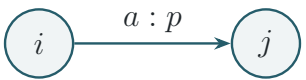
A **Mealy-type** (edge-emitting) HMM consists of:

- A finite set S of **hidden states**.
- A finite alphabet $\mathcal{X}$ of **observed symbols**.
- A collection of $|S| \times |S|$ **transition matrices** $(T^{(a)})_{a \in \mathcal{X}}$.
- An **initial distribution** $\eta^{(\emptyset)}$ over S.

The matrices specify the joint emission and state transition probabilities:

$$T_{ij}^{(a)} := \Pr(X = a,\ S' = j \mid S = i)\,.$$

We can conveniently represent these transitions with labelled diagrams:



$$p = T_{ij}^{(a)}$$

Moore-type HMMs emit from states; Mealy-type HMMs emit on transitions. They describe the same processes, but may require different numbers of hidden states.

Fix an initial distribution over hidden states, written as a row vector:

$$\eta_i^{(\emptyset)} := \Pr(S_0 = i), \qquad i \in S.$$

The probability of observing $x_{1:n} \in \mathcal{X}^n$ is given by:

$$\begin{aligned}\Pr(X_{1:n} = x_{1:n}) = \sum\nolimits_{i_0,\ldots,i_n \in S} & \Pr(S_0 = i_0) \\ & \times \Pr(X_1 = x_1, S_1 = i_1 \mid S_0 = i_0) \\ & \times \Pr(X_2 = x_2, S_2 = i_2 \mid S_1 = i_1) \cdots \\ & \times \Pr(X_n = x_n, S_n = i_n \mid S_{n-1} = i_{n-1}).\end{aligned}$$

Matrix multiplication performs these sums over hidden states:

$$\Pr(X_{1:n} = x_{1:n}) = \eta^{(\emptyset)} T^{(x_1)} T^{(x_2)} \cdots T^{(x_n)} \mathbf{1}^\top.$$

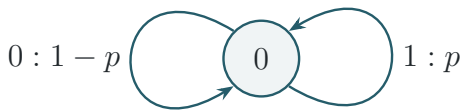
Here $\mathbf{1}$ is the $|S|$-dimensional row vector of ones.

Key point: probabilities are generated by a linear dynamic!

Example 1: Biased coin

Each observation is an independent coin toss, with $\Pr(X_t = 1) = p$, where $0 < p < 1$.

$$S = \{0\}, \qquad \mathcal{X} = \{0, 1\}, \qquad \eta^{(\emptyset)} = \begin{bmatrix} 1 \end{bmatrix}.$$



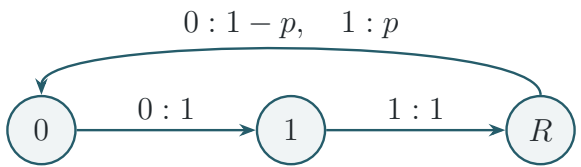
Example output: `1011001010`...

$$T^{(0)} = \begin{bmatrix} 1-p \end{bmatrix}, \qquad T^{(1)} = \begin{bmatrix} p \end{bmatrix}.$$

Example 2: Zero–One–Random

The generator cycles through three phases: emit 0, emit 1, then emit a random symbol with $\Pr(1) = p$, where $0 < p < 1$.

$$S = \{0, 1, R\}, \qquad \mathcal{X} = \{0, 1\}, \qquad \eta^{(\emptyset)} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$



Example output: `011 010 011 011` ... (starting in phase 0). State order: $(0, 1, R)$.

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

Predicting HMM observations

We have observed $X_{1:n} = x_{1:n}$ and want to predict the next symbol:

$$\Pr(X_{n+1} = a \mid X_{1:n} = x_{1:n}).$$

If we knew the current hidden state $S_n = i$, prediction would be straightforward:

$$\Pr(X_{n+1} = a \mid S_n = i) = \sum_{j \in S} T_{ij}^{(a)}.$$

With the hidden state unobserved, we average over its possible values:

$$\Pr(X_{n+1} = a \mid X_{1:n} = x_{1:n}) = \sum_{i \in S} \underbrace{\Pr(S_n = i \mid X_{1:n} = x_{1:n})}_{\text{belief state component}} \sum_{j \in S} T_{ij}^{(a)}.$$

The **belief state** is the row vector

$$\eta_i^{(x_{1:n})} := \Pr(S_n = i \mid X_{1:n} = x_{1:n}), \qquad i \in S$$

Computing and updating the belief state

For a history $x_{1:n}$ with positive probability, Bayes' rule gives

$$\eta_i^{(x_{1:n})} = \Pr(S_n = i \mid X_{1:n} = x_{1:n}) = \frac{\Pr(S_n = i, X_{1:n} = x_{1:n})}{\Pr(X_{1:n} = x_{1:n})}.$$

Using the HMM transition matrices,

$$\eta^{(x_{1:n})} = \frac{\eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)}}{\eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} \mathbf{1}^\top}.$$

When the next token x_{n+1} is observed, the belief updates recursively:

$$\eta^{(x_{1:n})} \xmapsto{x_{n+1}} \eta^{(x_{1:n+1})} = \frac{\eta^{(x_{1:n})} T^{(x_{n+1})}}{\eta^{(x_{1:n})} T^{(x_{n+1})} \mathbf{1}^\top}.$$

The current belief and the newly observed token determine the next belief. The full observation history does not need to be revisited.

The next-token probability vector

For a history $x_{1:n}$, define the row vector

$$p^{(x_{1:n})} := (\Pr(X_{n+1} = a \mid X_{1:n} = x_{1:n}))_{a \in \mathcal{X}}\,.$$

Returning to the expression on slide 9,

$$\Pr(X_{n+1} = a \mid X_{1:n} = x_{1:n}) = \sum_{i \in S} \Pr(S_n = i \mid X_{1:n} = x_{1:n}) \sum_{j \in S} T_{ij}^{(a)}.$$

Using the belief state and next-token vector, this becomes

$$p^{(x_{1:n})} = \eta^{(x_{1:n})} A, \qquad A_{ia} := \sum_{j \in S} T_{ij}^{(a)}.$$

The HMM fixes A, so the map from belief states to next-token vectors is linear. Distinct belief states can produce the same next-token vector.

Sufficient statistics for HMM prediction

Recall that a statistic R is **sufficient for prediction** if histories with the same statistic have the same probability for every finite continuation:

$$R(x) = R(x') \quad \Longrightarrow \quad \Pr(y \mid x) = \Pr(y \mid x') \qquad \text{for every } y \in \mathcal{X}^*.$$

The next-token probability vector $p^{(x_{1:n})}$ gives all one-symbol continuation probabilities, but not necessarily those for longer continuations. It need not be sufficient for the **entire future**. *(See exercises for details.)*

The belief state $\eta^{(x_{1:n})}$ preserves the probability of every finite continuation of an HMM. It is sufficient for predicting the **entire future**.

The belief state is a sufficient statistic

Given a possible history $x_{1:n}$ and finite continuation $y_{1:m}$, Bayes' rule gives

$$\begin{aligned}
&\Pr(X_{n+1:n+m} = y_{1:m} \mid X_{1:n} = x_{1:n}) \\
&\quad = \frac{\Pr(X_{1:n+m} = x_{1:n}y_{1:m})}{\Pr(X_{1:n} = x_{1:n})} \\
&\quad = \frac{\eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} T^{(y_1)} \cdots T^{(y_m)} \mathbf{1}^\top}{\eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} \mathbf{1}^\top} \\
&\quad = \eta^{(x_{1:n})} T^{(y_1)} \cdots T^{(y_m)} \mathbf{1}^\top.
\end{aligned}$$

The history enters this expression only through its belief state. Hence,

$$\eta^{(x)} = \eta^{(x')} \quad \Longrightarrow \quad \Pr(y \mid x) = \Pr(y \mid x') \qquad \text{for every } y \in \mathcal{X}^*.$$

The belief state is sufficient. **What about minimality?**

Belief states need not be minimal

Consider the two-state HMM

$$\eta^{(\emptyset)} = \left(\tfrac{1}{2}, \tfrac{1}{2}\right), \qquad T^{(0)} = \begin{pmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & 0 \end{pmatrix}, \qquad T^{(1)} = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{pmatrix}.$$

After observing 0 or 1, respectively, the belief state is

$$\eta^{(0)} = (1, 0), \qquad \eta^{(1)} = (0, 1).$$

However, the future is a sequence of independent fair coin flips in either case:

$$\Pr(y \mid 0) = \Pr(y \mid 1) = 2^{-|y|} \qquad \text{for every } y \in \{0,1\}^*.$$

Thus, distinct beliefs can contain exactly the same predictive information. More generally, define the **predictive equivalence relation**

$$\eta \sim \eta' \quad \iff \quad \eta T^{(y)} \mathbf{1}^\top = \eta' T^{(y)} \mathbf{1}^\top \qquad \text{for every } y \in \mathcal{X}^*.$$

Beliefs are **minimal** precisely when no two reachable beliefs are **predictively equivalent**. Generalised HMMs will let us characterise this condition.

Generalised hidden Markov models

A d-dimensional **generalised hidden Markov model** (GHMM) consists of:

- A finite alphabet $\mathcal{X}$ of **observed symbols**.
- A collection of $d \times d$ real **transition matrices** $(T^{(a)})_{a \in \mathcal{X}}$.
- An **initial row vector** $\eta^{(\emptyset)} \in \mathbb{R}^{1 \times d}$.
- A **column vector** $\phi \in \mathbb{R}^{d \times 1}$.

Writing $T = \sum_{a \in \mathcal{X}} T^{(a)}$, these objects must satisfy

$$T\phi = \phi, \qquad \eta^{(\emptyset)}\phi = 1, \qquad \eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} \phi \geq 0 \quad \text{for every } x_{1:n} \in \mathcal{X}^*.$$

These conditions ensure that the probability of any observed sequence is

$$\Pr(X_{1:n} = x_{1:n}) = \eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} \phi.$$

An HMM is the special case $d = |S|$ and $\phi = \mathbf{1}^\top$, where $\eta^{(\emptyset)}$ is a probability distribution, each $T^{(a)}$ is nonnegative, and T is row-stochastic.

Minimality of a GHMM

Histories and continuations generate two linear spaces:

$$\mathcal{H} := \operatorname{span}\Big\{ \eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} \;\Big|\; x_{1:n} \in \mathcal{X}^n,\ n \geq 0 \Big\},$$
$$\mathcal{F} := \operatorname{span}\Big\{ T^{(y_1)} \cdots T^{(y_m)} \phi \;\Big|\; y_{1:m} \in \mathcal{X}^m,\ m \geq 0 \Big\}.$$

A GHMM is **minimal-dimensional** if no lower-dimensional GHMM realises the same process. For a d-dimensional GHMM, this is equivalent to

$$\dim \mathcal{H} = \dim \mathcal{F} = d.$$

Theorem. If a stochastic process admits a GHMM realisation, then it admits a **minimal-dimensional GHMM realisation**.

Upper (1997), Theorem 4.3.10.

When HMM belief states are minimal

If an HMM is minimal-dimensional among all GHMM realisations of its process, then its belief state is a **minimal sufficient statistic** for prediction.

Suppose two supported histories $x_{1:n}$ and $x'_{1:n'}$ predict every continuation identically. Then, for every $y_{1:m}$,

$$\left(\eta^{(x_{1:n})} - \eta^{(x'_{1:n'})}\right) T^{(y_1)} \cdots T^{(y_m)} \mathbf{1}^\top = 0.$$

Minimal-dimensionality implies that the future vectors

$$T^{(y_1)} \cdots T^{(y_m)} \mathbf{1}^\top$$

span the entire column space. Therefore,

$$\eta^{(x_{1:n})} = \eta^{(x'_{1:n'})}.$$

Thus, histories have the same belief state precisely when they have the same distribution over the future. The belief state is **minimal sufficient**.

Based on Upper (1997), Sections 4.2–4.3.

Two notions of minimality

Here we summarise how each notion of minimality enters the key result.

- **Minimal sufficient statistics:** coarsest partition of the allowed sequences.
- **Minimal-dimensional:** no equivalent GHMM has lower dimension.

HMM is minimal-dimensional $\Longrightarrow$ belief states are minimal sufficient statistics.

Belief states (or their analogues) for a minimal-dimensional GHMM are the **lowest-dimensional** minimal sufficient statistics among those from which the entire future distribution is **linearly decodable**.

Working hypothesis: neural networks trained on GHMM data represent these objects. *(We will examine supporting evidence in the reading session.)*

Park, Choe & Veitch, *The Linear Representation Hypothesis and the Geometry of Large Language Models* (ICML 2024); Elhage et al., *Toy Models of Superposition* (Anthropic, 2022).

Summary

- A **belief state** is a sufficient statistic for prediction: the posterior distribution over an HMM's hidden states.
- The **next-token vector** need not be sufficient. It is a linear readout of the belief state, $p^{(x_{1:n})} = \eta^{(x_{1:n})} A$.
- GHMMs generalise HMMs by relaxing the entrywise probabilistic constraints on the transition matrices, while retaining the same matrix-product expression for output probabilities.
- Beliefs need not be minimal. If the HMM is **minimal-dimensional** as a GHMM, they are **minimal sufficient statistics**.