# Computational Mechanics

Xavier Poncini
Simplex

Adam Shai
Simplex

Paul Riechers
Simplex

## 1 Prerequisites

Familiarity with the goals and basic methodology of mechanistic interpretability, including the notions of features and circuits; the linear representation hypothesis; linear probes as a method for reading off internal representations; and a basic understanding of the transformer architecture.

Students should be comfortable with the idea that one can train a linear map from model activations to some target structure and evaluate its quality (e.g. via MSE or $R^2$).

**Mathematical background.** The module is mathematically self-contained: all formal definitions (HMMs, GHMMs, belief states, MSPs) are introduced from scratch. However, students will engage with the material more fluently if they are comfortable with the following:

- **Linear algebra:** row-stochastic matrices, rank and invertibility, null spaces (kernels).
- **Probability:** conditional probability, Bayes' rule, probability distributions over finite sets, the probability simplex.
- **Basic machine learning:** next-token prediction, loss functions (cross-entropy), the concept of model activations.

**What you'll learn**

- Construct their own hidden Markov models.
- Explain the motivation for introducing generalised hidden Markov models (GHMMs).
- Explain why the belief state is a useful object for making predictions about GHMM data.
- Compute the belief state of a GHMM.
- Explain the difference between the belief state of a HMM and the belief state of a GHMM (that is itself not an HMM).

- Explain the relationship between belief states and next-token observation probabilities for GHMMs.
- Compute the mixed state presentation (MSP) for GHMMs.
- Explain why generating data is easier than predicting data.
- Explain the evidence for the claim: transformers represent belief state geometry in their residual stream.
- Develop hypotheses for where transformers represent: belief states, observation probabilities, log observation probabilities.

# 2 Content

## 2.1 Fast-track

Read Transformers represent belief state geometry in their residual stream. Focus on being able to answer the following questions:

- What is a hidden Markov model (HMM)?
- What is a belief state?
- What is the mixed state presentation (MSP)?
- About the map learnt from activations to belief states:
  - What is the source?
  - What is the target?
  - How is the quality of the map evaluated?

## 2.2 Schedule

- **10:00–10:30** — Lecture: Overview and scope
- **10:30–11:15** — Lecture: Predicting the future
- **11:15–11:30** — *Break*
- **11:30–12:30** — Exercises: Predicting the future
- **12:30–13:30** — *Lunch*
- **13:30–14:15** — Lecture: Representing the past
- **14:15–15:45** — Exercises: Representing the past
- **15:45–16:00** — *Break*

- **16:00–17:45** — Reading & Discussion: Transformers represent belief geometry
- **17:45–18:00** — Feedback

## 3 Predicting the future

**Instructions**

Give a clear justification for each answer. You may use results proved in the lecture unless an exercise asks you to establish them. Questions marked **Extension** are optional.

**Instructions**

These notes reproduce each exercise before its solution. Equivalent arguments and notation should receive full credit when the mathematical reasoning is correct.

**Exercise 3.1 (From statistics to partitions).** Let $u_1, \dots, u_6 \in \mathcal{X}_+^*$ be six possible histories. Suppose that they induce the following conditional distributions over the entire future:

| history | $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ | $u_6$ |
|---|---|---|---|---|---|---|
| future distribution | $P$ | $P$ | $P$ | $Q$ | $Q$ | $S$ |

where $P$, $Q$, and $S$ are pairwise distinct.

Three statistics $R_A$, $R_B$, and $R_C$ induce the partitions

$$\Pi_A = \{\{u_1, u_2\}, \{u_3\}, \{u_4\}, \{u_5\}, \{u_6\}\},$$
$$\Pi_B = \{\{u_1, u_2, u_3\}, \{u_4, u_5\}, \{u_6\}\},$$
$$\Pi_C = \{\{u_1, u_2, u_3, u_4\}, \{u_5\}, \{u_6\}\}.$$

Recall that a statistic is sufficient for prediction when histories assigned the same value induce the same distribution over the future.

(a) What does it mean for two histories to belong to the same cell of the partition induced by a statistic? Illustrate your answer using $u_1$ and $u_2$ in $\Pi_A$.

(b) Determine which of $R_A$, $R_B$, and $R_C$ are sufficient. If a statistic is not sufficient, give two histories witnessing the failure.

(c) Which cells of $\Pi_A$ can be merged without losing sufficiency? Perform all such merges and identify the resulting partition.

(d) Explain why no two cells of $\Pi_B$ can be merged while preserving sufficiency. What does this show about $R_B$?

(e) Prove that every refinement of a sufficient partition is also sufficient. Is every coarsening of a sufficient partition necessarily sufficient? Use $\Pi_A$, $\Pi_B$, and $\Pi_C$ to illustrate your answer.

(f) **Extension.** Consider the partition

$$\Pi_D = \{\{u_1\}, \{u_2, u_3\}, \{u_4, u_5\}, \{u_6\}\}.$$

Show that $\Pi_A$ and $\Pi_D$ are both sufficient but that neither refines the other. How can both nevertheless refine the same minimal sufficient partition?

**Solution to Exercise 3.1**

(a) Two histories belong to the same cell when they receive the same summary value. In particular, $R_A(u_1) = R_A(u_2)$.

(b) Both $R_A$ and $R_B$ are sufficient. Every cell of $\Pi_A$ and every cell of $\Pi_B$ contains histories carrying only one of the distributions $P$, $Q$, or $S$.

The statistic $R_C$ is not sufficient. Its first cell contains $u_1$ and $u_4$, for example, but $u_1$ induces $P$ while $u_4$ induces $Q$. Since $P \neq Q$, these histories cannot be assigned the same value by a sufficient statistic.

(c) The cell $\{u_1, u_2\}$ can be merged with $\{u_3\}$ because all three histories induce $P$. Similarly, $\{u_4\}$ can be merged with $\{u_5\}$ because both histories induce $Q$. No cell can be merged with $\{u_6\}$, and no $P$-cell can be merged with a $Q$-cell. After all permissible merges, the resulting partition is

$$\{\{u_1, u_2, u_3\}, \{u_4, u_5\}, \{u_6\}\} = \Pi_B.$$

(d) The three cells of $\Pi_B$ correspond respectively to the three distinct future distributions $P$, $Q$, and $S$. Merging any two cells would therefore place histories with different future distributions in the same cell and destroy sufficiency. Hence $R_B$ is minimal sufficient.

(e) Let $\Pi'$ refine a sufficient partition $\Pi$. Any two histories in the same cell of $\Pi'$ also belong to the same cell of $\Pi$. Since $\Pi$ is sufficient, those histories induce the same future distribution. Therefore $\Pi'$ is sufficient.

In the example, $\Pi_A$ refines $\Pi_B$, and both are sufficient. A coarsening need not remain sufficient: $\Pi_C$ is obtained from the sufficient partition $\Pi_A$ by merging its cells $\{u_1, u_2\}$, $\{u_3\}$, and $\{u_4\}$. The resulting cell contains histories inducing both $P$ and $Q$, so $\Pi_C$ is not sufficient.

(f) Every cell of $\Pi_D$ lies within one predictive class, so $\Pi_D$ is sufficient. The same was established for $\Pi_A$.

The cell $\{u_1, u_2\}$ of $\Pi_A$ is split between $\{u_1\}$ and $\{u_2, u_3\}$ in $\Pi_D$, so $\Pi_A$ does not refine $\Pi_D$. Conversely, $\{u_2, u_3\}$ in $\Pi_D$ is split between $\{u_1, u_2\}$ and $\{u_3\}$ in $\Pi_A$, so $\Pi_D$ does not refine $\Pi_A$. Both partitions nevertheless refine $\Pi_B$, because every one of their cells is contained in one of the three predictive classes represented by $P$, $Q$, and $S$.

**Exercise 3.2 (Causal states are minimal sufficient statistics).** For $x, x' \in \mathcal{X}_+^*$, define the relation $\sim$ by

$$x \sim x' \iff \Pr\bigl(X_{|x|+1:|x|+|y|} = y \mid X_{1:|x|} = x\bigr) = \Pr\bigl(X_{|x'|+1:|x'|+|y|} = y \mid X_{1:|x'|} = x'\bigr)$$

for every $y \in \mathcal{X}^*$. Define the causal state of $x$ by

$$\epsilon(x) := [x]_\sim,$$

and write $\mathcal{S} := \epsilon(\mathcal{X}_+^*)$ for the set of causal states.

(a) Prove that $\sim$ is an equivalence relation on $\mathcal{X}_+^*$ by establishing each of the following:

  (i) reflexivity: $x \sim x$ for every $x \in \mathcal{X}_+^*$;

  (ii) symmetry: if $x \sim x'$, then $x' \sim x$;

  (iii) transitivity: if $x \sim x'$ and $x' \sim x''$, then $x \sim x''$.

(b) Prove that $\epsilon$ is sufficient for prediction.

(c) Let $R : \mathcal{X}_+^* \to \mathcal{R}$ be any sufficient statistic. Prove that

$$R(x) = R(x') \implies \epsilon(x) = \epsilon(x').$$

Deduce that every sufficient partition refines the causal-state partition.

(d) Using parts (b) and (c), conclude that $\epsilon$ is a minimal sufficient statistic.

(e) **Extension.** Let $T$ be another minimal sufficient statistic. Prove that $T$ and $\epsilon$ induce the same partition of $\mathcal{X}_+^*$. Deduce that there is a bijection

$$\psi : \mathcal{S} \to \mathrm{im}(T)$$

satisfying $T = \psi \circ \epsilon$.

**Solution to Exercise 3.2**

(a) (i) For every $x \in \mathcal{X}_+^*$ and every continuation $y$, the conditional probability after $x$ equals itself. Hence $x \sim x$, so $\sim$ is reflexive.

(ii) If $x \sim x'$, then the two continuation probabilities are equal for every $y$. Reversing the equality shows that $x' \sim x$, so $\sim$ is symmetric.

(iii) Suppose $x \sim x'$ and $x' \sim x''$. For every $y$, the continuation probability after $x$ equals that after $x'$, which equals that after $x''$. Thus $x \sim x''$, so $\sim$ is transitive.

Therefore $\sim$ is an equivalence relation.

(b) If $\epsilon(x) = \epsilon(x')$, then $x$ and $x'$ belong to the same equivalence class, so $x \sim x'$. By the definition of $\sim$, the two histories induce the same probability for every finite continuation $y$. This is precisely the definition of sufficiency for prediction, so $\epsilon$ is sufficient.

(c) Suppose $R(x) = R(x')$. Since $R$ is sufficient, $x$ and $x'$ induce the same probability for every finite continuation. Hence $x \sim x'$, and equivalent histories have the same equivalence class: $\epsilon(x) = \epsilon(x')$.

Thus any two histories in the same cell induced by $R$ also lie in the same causal state. Every cell induced by $R$ is therefore contained in a causal state, so the partition induced by $R$ refines the causal-state partition. Since $R$ was arbitrary, every sufficient partition has this property.

(d) Part (b) shows that $\epsilon$ is sufficient. Part (c) shows that its partition is refined by the partition of every sufficient statistic. Hence the causal-state partition is the coarsest sufficient partition, so $\epsilon$ is a minimal sufficient statistic.

(e) Since $T$ is sufficient, part (c), with $R = T$, shows that the partition induced by $T$ refines the causal-state partition. Since $T$ is minimal and $\epsilon$ is sufficient, the causal-state partition must also refine the partition induced by $T$. The two partitions are therefore equal.

Define $\psi(\epsilon(x)) := T(x)$. Equality of the two partitions shows that this definition is independent of the representative $x$ and that distinct causal states receive distinct $T$-values. It also reaches every value in $\mathrm{im}(T)$. Hence $\psi$ is a bijection.

**Exercise 3.3 (From finite generators to causal states).** In this exercise, $S_t$ denotes the hidden state immediately before the transition that emits $X_{t+1}$. All parameters lie strictly between 0 and 1.

(a) **The Zero–One–Random Process.** Let $S_0$ be uniformly distributed over

$\{s_0, s_1, s_R\}$. The process has transitions

$$s_0 \xrightarrow{0:1} s_1, \qquad s_1 \xrightarrow{1:1} s_R, \qquad s_R \begin{cases} \xrightarrow{0:1-p} s_0, \\ \xrightarrow{1:p} s_0. \end{cases}$$

An edge label $x : q$ means that the process emits $x$ and follows the edge with probability $q$.

(i) List all possible histories of length two. For each history, determine whether it uniquely identifies the current hidden phase. When it does, identify that phase.

(ii) For each ambiguous length-two history, determine which additional observations resolve the ambiguity. Identify the resulting phase and explain why it remains known after every subsequent observation.

(iii) Assume that $\epsilon(\emptyset)$, $\epsilon(0)$, $\epsilon(1)$, and $\epsilon(10)$ are pairwise distinct. Show that the causal states have shortest-length representatives

$$\emptyset, \quad 0, \quad 1, \quad 10, \quad 00, \quad 01, \quad 11.$$

Explain why no additional causal states arise.

(iv) Under the assumption in part (iii), draw the automaton associated with the causal-state dynamics.

(b) **The Even Process.** Let $S_0$ be uniformly distributed over $\{A, B\}$ and consider the generator

$$A \xrightarrow{0:q} A, \qquad A \xrightarrow{1:1-q} B, \qquad B \xrightarrow{1:1} A.$$

Thus every run of 1s between successive 0s has even length.

(i) For each of the length-one histories 0 and 1, determine whether the current hidden state can be identified exactly.

(ii) Starting from each ambiguous history in part (i), consider its possible one-symbol extensions. Which extensions resolve the ambiguity, and which remain ambiguous? You may use the fact that $\epsilon(\emptyset) = \epsilon(11)$. What does this imply for histories of the form $1^n$?

(iii) Assume additionally that $\epsilon(\emptyset) \neq \epsilon(1)$. Show that the causal states have shortest representatives

$$\emptyset, \qquad 1, \qquad 0, \qquad 01,$$

and explain why there are no additional causal states. You may distinguish causal states using possible and impossible future continuations.

(iv) Using $\epsilon(\emptyset) = \epsilon(11)$ and assuming $\epsilon(\emptyset) \neq \epsilon(1)$, draw the automaton associated with the causal-state dynamics.

(c) **The Simple Nonunifilar Source.** Let $S_0$ be uniformly distributed over $\{A, B\}$ and consider

$$A \xrightarrow{0:1/2} A, \qquad A \xrightarrow{0:1/2} B, \qquad B \xrightarrow{0:1/2} B, \qquad B \xrightarrow{1:1/2} A.$$

(i) Show that observing a 1 identifies the new hidden state as $A$.

(ii) Starting from $A$, show that there are $n+1$ hidden paths that emit $0^n$, each with probability $2^{-n}$. How many of these paths can subsequently emit 1? Hence calculate

$$\Pr(X_{n+2} = 1 \mid X_{1:n+1} = 10^n).$$

(iii) Show that this probability is different for every $n \geq 0$. Conclude that the histories

$$1, 10, 100, \ldots$$

belong to distinct causal states and that the process has infinitely many causal states.

**Solution to Exercise 3.3**

(a) (i) All four length-two histories are possible. Their resulting hidden phases are

| $x_{1:2}$ | 00 | 01 | 10 | 11 |
|---|---|---|---|---|
| $S_2$ | $s_1$ | $s_R$ | $\{s_0, s_1\}$ | $s_0$. |

Thus 00, 01, and 11 identify the phase exactly, while 10 is ambiguous.

(ii) From the phases compatible with 10, only $s_0$ can emit 0, and only $s_1$ can emit 1. Consequently, 100 identifies the new phase as $s_1$, while 101 identifies it as $s_R$. From each phase, the emitted symbol determines the next phase, so every later observation preserves synchronization.

(iii) The histories 00, 01, and 11 identify $s_1$, $s_R$, and $s_0$, respectively. These phases induce distinct next-token distributions: $s_1$ emits only 1, $s_0$ emits only 0, and $s_R$ can emit either symbol. Hence they give three distinct causal states.

They are also distinct from the four states assumed in the question. This can be seen without calculating probabilities by comparing possible length-two continuations:

| representative | possible length-two continuations |
|---|---|
| $\emptyset$ | $\{00, 01, 10, 11\}$ |
| 0 | $\{01, 10, 11\}$ |
| 1 | $\{00, 01, 10\}$ |
| 10 | $\{01, 10, 11\}$ |
| 00 | $\{10, 11\}$ |
| 01 | $\{00, 10\}$ |
| 11 | $\{01\}$ |

The only repeated continuation set belongs to 0 and 10, whose causal states are assumed to be distinct. Finally, every history of length at least three identifies the phase by part (ii), and so belongs to one of the states represented by 00, 01, and 11. No further causal states arise.

(iv) Appending each possible symbol gives the causal-state update:

| state | 0 | 1 |
|---|---|---|
| $\epsilon(\emptyset)$ | $\epsilon(0)$ | $\epsilon(1)$ |
| $\epsilon(0)$ | $\epsilon(00)$ | $\epsilon(01)$ |
| $\epsilon(1)$ | $\epsilon(10)$ | $\epsilon(11)$ |
| $\epsilon(10)$ | $\epsilon(00)$ | $\epsilon(01)$ |
| $\epsilon(00)$ | – | $\epsilon(01)$ |
| $\epsilon(01)$ | $\epsilon(11)$ | $\epsilon(11)$ |
| $\epsilon(11)$ | $\epsilon(00)$ | – |

Here a dash denotes an impossible extension. Drawing one node for each row and the indicated symbol-labelled edges gives the required automaton.

(b) (i) The history 0 identifies the current state as $A$: only $A$ can emit 0, and this transition returns to $A$. The history 1 is ambiguous. If $S_0 = A$, it leads to $B$, while if $S_0 = B$, it leads to $A$.

(ii) From the two states compatible with the history 1, only $A$ can emit 0. Therefore 10 resolves the ambiguity and identifies the new state as $A$. The extension 11 remains ambiguous. Using the given equality and recursively updating after pairs of 1s gives

$$\epsilon(1^{2k}) = \epsilon(\emptyset), \qquad \epsilon(1^{2k+1}) = \epsilon(1), \qquad k \geq 0.$$

(iii) If a history contains a 0, its final 0 synchronizes the generator to $A$. An even number of trailing 1s then leaves it in $A$, represented by 0, while an odd number leaves it in $B$, represented by 01. If the history contains no 0, part (ii) shows that its causal state is represented by $\emptyset$ or 1, according to the parity of its length. Thus no additional causal states arise.

These four states are distinct. The state represented by 01 cannot emit 0, whereas the other three can. The continuation 10 is impossible after 0 but possible after $\emptyset$ and 1. Finally, $\epsilon(\emptyset) \neq \epsilon(1)$ by assumption. Hence the displayed representatives are also shortest.

(iv) Appending each possible symbol gives the causal-state update:

| state | 0 | 1 |
|---|---|---|
| $\epsilon(\emptyset)$ | $\epsilon(0)$ | $\epsilon(1)$ |
| $\epsilon(1)$ | $\epsilon(0)$ | $\epsilon(\emptyset)$ |
| $\epsilon(0)$ | $\epsilon(0)$ | $\epsilon(01)$ |
| $\epsilon(01)$ | – | $\epsilon(0)$ |

Here a dash denotes an impossible extension. Drawing one node for each row and the indicated symbol-labelled edges gives the required automaton.

(c) (i) The only transition that emits 1 starts from $B$, and that transition leads to $A$. Thus the hidden state immediately after observing 1 is $A$.

(ii) One path remains in $A$ while emitting all $n$ zeros. Each of the other $n$ paths moves from $A$ to $B$ on one of the $n$ transitions and then remains in $B$. Thus there are $n+1$ paths, each with probability $2^{-n}$, and

$$\Pr(0^n \mid S = A) = (n+1)2^{-n}.$$

The $n$ paths that end in $B$ can subsequently emit 1. Including this final transition, each has probability $2^{-(n+1)}$, so

$$\Pr(X_{n+2} = 1 \mid X_{1:n+1} = 10^n) = \frac{n2^{-(n+1)}}{(n+1)2^{-n}} = \frac{n}{2(n+1)}.$$

(iii) Since

$$\frac{n}{2(n+1)} = \frac{1}{2} - \frac{1}{2(n+1)},$$

this probability is strictly increasing with $n$. Hence every pair of histories $10^n$ and $10^m$, with $n \neq m$, already has a different next-token distribution. They must lie in different causal states. The source consequently has infinitely many causal states despite its two-state generator.

**Exercise 3.4 (Why next-token prediction is enough).** Recall that

$$\mathcal{X}_+^* := \big\{x \in \mathcal{X}^* : \Pr\big(X_{1:|x|} = x\big) > 0\big\}$$

is the set of allowable histories.
Consider a model with parameters $\theta$ satisfying

$$h_n = f_\theta(h_{n-1}, x_n), \qquad y_n = g_\theta(h_{n-1}, x_n).$$

Fix the initial hidden state $h_0$ and suppose that the model predicts the next-token distribution exactly: for every $n \geq 1$,

$$y_n = \Pr(X_{n+1} \mid X_{1:n} = x_{1:n}), \qquad \forall\, x_{1:n} \in \mathcal{X}_+^*.$$

For every nonempty allowable history, define

$$R(x_{1:n}) := (h_{n-1}, x_n).$$

For $a \in \mathcal{X}$, define the deterministic update

$$\Phi_a(h, x) := \big(f_\theta(h, x), a\big).$$

For a word $z_{1:j}$, write

$$\Phi_{z_{1:j}} := \Phi_{z_j} \circ \cdots \circ \Phi_{z_1}, \qquad \Phi_{z_{1:0}} := \mathrm{id}\,.$$

(a) Show that, whenever $x_{1:n}z_{1:j}$ is allowable,

$$R(x_{1:n}z_{1:j}) = \Phi_{z_{1:j}}(R(x_{1:n}))\,.$$

Hence show that, if $R(x_{1:n}) = R(x'_{1:m})$ and both extended histories are allowable, then

$$R(x_{1:n}z_{1:j}) = R(x'_{1:m}z_{1:j}).$$

(b) Let $R(x_{1:n}) = R(x'_{1:m})$ and fix $z_{1:k} \in \mathcal{X}^k$. As the symbols of $z_{1:k}$ are read from left to right, use part (a) and exact next-token prediction to show that for each $0 \leq j < k$ such that the prefixes through $z_{1:j}$ are allowable,

$$\begin{aligned}&\Pr(X_{n+j+1} = z_{j+1} \mid X_{1:n+j} = x_{1:n}z_{1:j})\\ &\qquad = \Pr\big(X_{m+j+1} = z_{j+1} \mid X_{1:m+j} = x'_{1:m}z_{1:j}\big)\,.\end{aligned}$$

Deduce that either the same first zero-probability extension is encountered after both histories, or every prefix of $z_{1:k}$ is allowable after both histories.

(c) In the case where every prefix is allowable, use the conditional chain rule to write

$$\begin{aligned}&\Pr(X_{n+1:n+k} = z_{1:k} \mid X_{1:n} = x_{1:n}) \\ &\quad = \prod_{j=0}^{k-1} \Pr(X_{n+j+1} = z_{j+1} \mid X_{1:n+j} = x_{1:n} z_{1:j})\,.\end{aligned}$$

Write the analogous product after $x'_{1:m}$ and compare its factors. Deal separately with the zero-probability case from part (b), and conclude directly that

$$R(x_{1:n}) = R(x'_{1:m}) \quad \Longrightarrow$$
$$\Pr(X_{n+1:n+k} = z_{1:k} \mid X_{1:n} = x_{1:n}) = \Pr\big(X_{m+1:m+k} = z_{1:k} \mid X_{1:m} = x'_{1:m}\big)\,.$$

Hence $R(x_{1:n}) = (h_{n-1}, x_n)$ is a sufficient statistic for predicting the entire future.

**Solution to Exercise 3.4**

(a) Appending one symbol $a$ gives

$$R(x_{1:n}a) = \big(f_\theta(h_{n-1}, x_n), a\big) = \Phi_a(R(x_{1:n}))\,.$$

Repeatedly applying this deterministic update along $z_{1:j}$ gives

$$R(x_{1:n} z_{1:j}) = \Phi_{z_j} \circ \cdots \circ \Phi_{z_1}(R(x_{1:n})) = \Phi_{z_{1:j}}(R(x_{1:n}))\,.$$

If $R(x_{1:n}) = R(x'_{1:m})$ and both extended histories are allowable, applying the same composite function to the equal starting representations gives

$$R(x_{1:n} z_{1:j}) = R(x'_{1:m} z_{1:j}).$$

(b) Suppose that the prefixes through $z_{1:j}$ are allowable after both histories. Part (a) gives

$$R(x_{1:n} z_{1:j}) = R(x'_{1:m} z_{1:j}).$$

The two representations give the same arguments to $g_\theta$, so exact next-token prediction gives

$$\begin{aligned}&\Pr(X_{n+j+1} = z_{j+1} \mid X_{1:n+j} = x_{1:n} z_{1:j}) \\ &\quad = \Pr\big(X_{m+j+1} = z_{j+1} \mid X_{1:m+j} = x'_{1:m} z_{1:j}\big)\,.\end{aligned}$$

Thus the next extension has positive probability after one history exactly when it has positive probability after the other. Following the word from left to right, either both encounter their first zero at the same position or all its prefixes are allowable after both.

(c) If every prefix is allowable, the conditional chain rule gives the product in the question and

$$\begin{aligned}&\Pr\bigl(X_{m+1:m+k} = z_{1:k} \mid X_{1:m} = x'_{1:m}\bigr)\\&\quad = \prod_{j=0}^{k-1} \Pr\bigl(X_{m+j+1} = z_{j+1} \mid X_{1:m+j} = x'_{1:m} z_{1:j}\bigr)\,.\end{aligned}$$

Part (b) shows that the factors in the two products are equal, so the products are equal. If a first zero-probability extension is encountered instead, part (b) shows that it occurs after both histories; consequently both probabilities of the complete continuation are zero.

Therefore equal values of $R$ imply equal conditional probabilities for every finite continuation. By the definition of sufficiency,

$$R(x_{1:n}) = (h_{n-1}, x_n)$$

is a sufficient statistic for predicting the entire future.

# 4 Representing the past

**Instructions**

Give a clear justification for each answer. You may use results proved in the lecture unless an exercise asks you to establish them. Questions marked **Extension** are optional.

**Instructions**

These notes reproduce each exercise before its solution. Equivalent arguments, diagrams, and notation should receive full credit when the mathematical reasoning is correct.

**Exercise 4.1 (Translating between HMM diagrams and matrices).** For an edge-emitting hidden Markov model, use the convention

$$T_{ij}^{(a)} := \Pr(X_{t+1} = a, S_{t+1} = j \mid S_t = i).$$

Thus rows index the current hidden state, columns index the next hidden state, and the superscript records the emitted symbol. Throughout this exercise, $0 < p, q < 1$.

(a) Suppose that $T_{ij}^{(a)} = r$. Explain each component of the following diagram:

$$i \xrightarrow{a:r} j.$$

What does a zero entry mean? Why is the matrix $T := \sum_{a\in\mathcal{X}} T^{(a)}$ row-stochastic?

(b) **From matrices to a diagram: Zero–One–Random.** Take the state order to be $(0, 1, R)$ and suppose that

$$\eta^{(\emptyset)} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}, \qquad T^{(0)} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1-p & 0 & 0 \end{bmatrix}, \qquad T^{(1)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ p & 0 & 0 \end{bmatrix}.$$

(i) Draw the corresponding edge-emitting HMM. Describe informally the three phases represented by 0, 1, and $R$.

(ii) Calculate $\Pr(X_{1:3} = 011)$ and $\Pr(X_{1:3} = 000)$ in two ways: first by tracing paths through your diagram, and then from

$$\Pr(X_{1:3} = x_{1:3}) = \eta^{(\emptyset)} T^{(x_1)} T^{(x_2)} T^{(x_3)} \mathbf{1}.$$

(c) **From a diagram to matrices: the Even Process.** Consider the edge-emitting HMM

$$A \xrightarrow{0:q} A, \qquad A \xrightarrow{1:1-q} B, \qquad B \xrightarrow{1:1} A,$$

with no other edges. Take the state order to be $(A, B)$ and let $\eta^{(\emptyset)} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}$.

(i) Construct $T^{(0)}$ and $T^{(1)}$. Account explicitly for every zero entry in the two matrices.

(ii) Form $T = T^{(0)} + T^{(1)}$ and verify that it is row-stochastic.

(iii) Calculate $\Pr(X_{1:3} = 011)$ and $\Pr(X_{1:3} = 010)$ both by tracing paths through the diagram and by multiplying the appropriate symbol matrices. Explain how the second calculation reflects the defining constraint of the Even Process.

(d) **Extension.** Let $\mathcal{X}$ be the observation alphabet of an arbitrary finite-state HMM, let $T = \sum_{a\in\mathcal{X}} T^{(a)}$, and let $\eta^{(\emptyset)}$ be a probability distribution. Prove that, for every $n \geq 0$,

$$\sum_{w\in\mathcal{X}^n} \Pr(X_{1:n} = w) = \eta^{(\emptyset)} T^n \mathbf{1} = 1.$$

**Solution to Exercise 4.1**

(a) In the diagram, $i$ is the current hidden state, $j$ is the next hidden state, $a$ is the emitted symbol, and $r$ is the probability of that joint emission-transition event conditional on starting in $i$. Thus the edge records $T_{ij}^{(a)} = \Pr(X_{t+1} = a, S_{t+1} = j \mid S_t = i) = r$. A zero entry means that no such emission-transition pair is possible. For a fixed current state $i$, the events indexed by all pairs $(a, j)$ exhaust the possible emitted symbols and next states. Consequently,

$$\sum_{a\in\mathcal{X}} \sum_{j} T_{ij}^{(a)} = 1.$$

Since

$$T_{ij} = \sum_{a\in\mathcal{X}} T_{ij}^{(a)},$$

every entry of $T$ is nonnegative and its $i$th row sums to

$$\sum_{j} T_{ij} = \sum_{j} \sum_{a\in\mathcal{X}} T_{ij}^{(a)} = 1.$$

Hence $T$ is row-stochastic.

(b) (i) The nonzero entries give the transitions

$$0 \xrightarrow{0:1} 1, \qquad 1 \xrightarrow{1:1} R, \qquad R \xrightarrow{0:1-p} 0, \qquad R \xrightarrow{1:p} 0.$$

Thus the diagram is a three-state cycle, with two differently labelled edges from $R$ to 0. State 0 is the phase that emits 0 deterministically, state 1 is the phase that emits 1 deterministically, and state $R$ is the random phase, which emits 1 with probability $p$ and 0 with probability $1 - p$.

(ii) For 011, the only possible initial state is 0. The model then follows

$$0 \xrightarrow{0:1} 1 \xrightarrow{1:1} R \xrightarrow{1:p} 0.$$

Including the initial probability $1/3$ gives $\Pr(011) = p/3$. Matrix multiplication gives the same result:

$$\eta^{(\emptyset)} T^{(0)} T^{(1)} T^{(1)} = \begin{bmatrix} p/3 & 0 & 0 \end{bmatrix}, \qquad \Pr(011) = p/3.$$

No state permits a path labelled 000: starting from 0, the first 0 moves to state 1, which cannot emit 0; starting from 1, even the first 0 is impossible; and starting from $R$, the first two zeros move through $R \to 0 \to 1$, after which a third zero is impossible. Correspondingly,

$$\eta^{(\emptyset)} T^{(0)} T^{(0)} T^{(0)} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, \qquad \Pr(000) = 0.$$

(c) (i) In state order $(A, B)$, the matrices are

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

In $T^{(0)}$, the entry $(A, B)$ is zero because no 0-edge connects $A$ to $B$, while the entire $B$ row is zero because state $B$ cannot emit 0. In $T^{(1)}$, the diagonal entries are zero because neither 1-edge is a self-loop. The two off-diagonal entries record $A \xrightarrow{1:1-q} B$ and $B \xrightarrow{1:1} A$.

(ii) We have

$$T = \begin{bmatrix} q & 1-q \\ 1 & 0 \end{bmatrix}.$$

Its entries are nonnegative and both rows sum to one.

(iii) The word 011 can be produced only by starting from $A$ and following

$$A \xrightarrow{0:q} A \xrightarrow{1:1-q} B \xrightarrow{1:1} A.$$

Thus

$$\Pr(011) = \frac{1}{2}q(1-q).$$

The matrix calculation is

$$\eta^{(\emptyset)}T^{(0)}T^{(1)}T^{(1)} = \begin{bmatrix} \frac{1}{2}q(1-q) & 0 \end{bmatrix}, \qquad \Pr(011) = \frac{1}{2}q(1-q).$$

The word 010 would require the path to emit a 0 immediately after the transition $A \xrightarrow{1} B$, but state $B$ cannot emit 0. Hence there is no such path. Equivalently,

$$\eta^{(\emptyset)}T^{(0)}T^{(1)}T^{(0)} = \begin{bmatrix} 0 & 0 \end{bmatrix}, \qquad \Pr(010) = 0.$$

This is the shortest example of an odd run of 1s occurring between two 0s, which the Even Process forbids.

(d) Expanding the product of the sum of the symbol matrices gives

$$T^n = \left( \sum_{a \in \mathcal{X}} T^{(a)} \right)^n = \sum_{w \in \mathcal{X}^n} T^{(w_1)} \cdots T^{(w_n)}.$$

Multiplying on the left by $\eta^{(\emptyset)}$ and on the right by $\mathbf{1}$ therefore yields

$$\eta^{(\emptyset)}T^n\mathbf{1} = \sum_{w \in \mathcal{X}^n} \eta^{(\emptyset)}T^{(w_1)} \cdots T^{(w_n)}\mathbf{1} = \sum_{w \in \mathcal{X}^n} \Pr(w).$$

Since $T$ is row-stochastic, $T\mathbf{1} = \mathbf{1}$ and hence $T^n\mathbf{1} = \mathbf{1}$. Finally, $\eta^{(\emptyset)}\mathbf{1} = 1$, because the initial vector is a probability distribution. The required sum is therefore one.

**Exercise 4.2 (Belief states and a lossy next-token readout).** Continue with the Zero–One–Random HMM from Exercise 4.1, using state order $(0, 1, R)$. For an allowable history $x_{1:n} = x_1 \cdots x_n$, compute its belief from scratch as

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

Here $\eta_i^{(x_{1:n})} = \Pr(S_n = i \mid X_{1:n} = x_{1:n})$. For an allowable extension $x_{1:n+1} = x_{1:n}a$, update recursively as

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

(a) Compute $\eta^{(0)}$ and $\eta^{(1)}$. Interpret their components as posterior probabilities over the three hidden phases.

(b) Show that the seven beliefs

$$\mathcal{B} = \{\eta^{(\emptyset)}, \eta^{(0)}, \eta^{(1)}, \eta^{(00)}, \eta^{(01)}, \eta^{(10)}, \eta^{(11)}\}$$

are distinct and that $\mathcal{B}$ is closed under every allowable symbol update. Draw the symbol-labelled transition diagram on $\mathcal{B}$.

(c) Compare your belief-state diagram with both the causal-state automaton constructed in Part 2 and the three-state generative HMM from Exercise 4.1.

(i) Match each belief to the causal state with the same shortest-length representative. Do the symbol-labelled transitions agree?

(ii) Which beliefs correspond to knowing the hidden phase exactly, and which represent uncertainty over several phases? Show that the hidden phase becomes known after at most three observations and remains known thereafter.

(iii) Observe that the finite-history belief-state presentation has seven states while the generative HMM has only three. What does this suggest about the possible relative difficulty of generation and inference?

(d) For each belief $\eta^{(x_{1:n})} \in \mathcal{B}$ from part (b), compute its associated next-token probability vector

$$p^{(x_{1:n})} := (\Pr(X_{n+1} = a \mid X_{1:n} = x_{1:n}))_{a \in \{0,1\}} = \eta^{(x_{1:n})} A,$$

$$A_{ia} := \sum_j T_{ij}^{(a)}.$$

Visualise the belief geometry on the 2-simplex over hidden states $(0, 1, R)$ and the next-token geometry on the 1-simplex over symbols $(0, 1)$.

(e) Calculate

$$\Pr(X_{3:4} = 00 \mid X_{1:2} = 10) \qquad \text{and} \qquad \Pr(X_{3:4} = 00 \mid X_{1:2} = 01).$$

Deduce that the next-token probability vector need not be sufficient for predicting the entire future.

(f) **Extension.** Let $\delta := \eta^{(10)} - \eta^{(01)}$. Show that

$$\delta A = 0 \qquad \text{but} \qquad \delta T^{(0)} T^{(0)} \mathbf{1} \neq 0.$$

Interpret these two statements in terms of one-step and longer-horizon prediction.

**Solution to Exercise 4.2**

(a) Multiplying the initial belief by the two symbol matrices gives

$$\eta^{(\emptyset)} T^{(0)} = \begin{bmatrix} (1-p)/3 & 1/3 & 0 \end{bmatrix}, \qquad \eta^{(\emptyset)} T^{(1)} = \begin{bmatrix} p/3 & 0 & 1/3 \end{bmatrix}.$$

Their row sums are respectively $(2-p)/3$ and $(1+p)/3$. Normalising therefore gives

$$\eta^{(0)} = \begin{bmatrix} \dfrac{1-p}{2-p} & \dfrac{1}{2-p} & 0 \end{bmatrix}, \qquad \eta^{(1)} = \begin{bmatrix} \dfrac{p}{1+p} & 0 & \dfrac{1}{1+p} \end{bmatrix}.$$

After observing 0, the current phase is either 0 or 1 with the displayed posterior probabilities, and cannot be $R$. After observing 1, it is either 0 or $R$, and cannot be 1.

(b) Direct evaluation gives

$$\begin{aligned} \eta^{(00)} &= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}, & \eta^{(01)} &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}, \\ \eta^{(10)} &= \begin{bmatrix} 1-p & p & 0 \end{bmatrix}, & \eta^{(11)} &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Thus the pure beliefs are $\eta^{(11)}$, $\eta^{(00)}$, and $\eta^{(01)}$. The pure beliefs are distinct. The beliefs $\eta^{(\emptyset)}$, $\eta^{(0)}$, and $\eta^{(1)}$ have different supports from one another and from the pure beliefs. The only remaining comparison is between $\eta^{(0)}$ and $\eta^{(10)}$, which both have support $\{0, 1\}$. Equality would require $1/(2-p) = p$, or $(1-p)^2 = 0$, contrary to $p < 1$.

The allowable updates are

| current belief | 0 | 1 |
|---|---|---|
| $\eta^{(\emptyset)}$ | $\eta^{(0)}$ | $\eta^{(1)}$ |
| $\eta^{(0)}$ | $\eta^{(00)}$ | $\eta^{(01)}$ |
| $\eta^{(1)}$ | $\eta^{(10)}$ | $\eta^{(11)}$ |
| $\eta^{(10)}$ | $\eta^{(00)}$ | $\eta^{(01)}$ |
| $\eta^{(11)}$ | $\eta^{(00)}$ | impossible |
| $\eta^{(00)}$ | impossible | $\eta^{(01)}$ |
| $\eta^{(01)}$ | $\eta^{(11)}$ | $\eta^{(11)}$ |

This table specifies the requested symbol-labelled diagram and shows that every allowable update stays in $\mathcal{B}$.

(c) (i) Under the assumptions used in Part 2, the correspondence is

$$\begin{aligned}&\epsilon(\emptyset) \longleftrightarrow \eta^{(\emptyset)}, \quad \epsilon(0) \longleftrightarrow \eta^{(0)}, \quad \epsilon(1) \longleftrightarrow \eta^{(1)},\\ &\epsilon(10) \longleftrightarrow \eta^{(10)}, \quad \epsilon(11) \longleftrightarrow \eta^{(11)}, \quad \epsilon(00) \longleftrightarrow \eta^{(00)},\\ &\epsilon(01) \longleftrightarrow \eta^{(01)}.\end{aligned}$$

The transition table from part (b) agrees with the causal-state automaton after making these substitutions.

(ii) The pure beliefs $\eta^{(11)}$, $\eta^{(00)}$, and $\eta^{(01)}$ identify the hidden phase exactly. The other four beliefs assign positive probability to more than one phase and therefore represent uncertainty. After two observations only the history 10 leaves a mixed belief. Either allowable third symbol sends $\eta^{(10)}$ to a pure belief. Once the belief is pure, every allowable transition in the table leads to another pure belief, so the phase remains known.

(iii) The generator has access to its actual hidden phase and needs only the three states 0, 1, and $R$. An observer begins with an unknown phase and must also represent four transient posterior distributions before synchronising with the generator. This example therefore shows that inference can require a richer state description than a particular generative presentation.

(d) The probability of each symbol from a hidden phase is obtained by summing the corresponding row of its symbol matrix over destination states. With column order $(0, 1)$, this gives

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1-p & p \end{bmatrix}.$$

Hence

| belief | next-token vector |
|---|---|
| $\eta^{(\emptyset)}$ | $p^{(\emptyset)} = \left(\dfrac{2-p}{3}, \dfrac{1+p}{3}\right)$ |
| $\eta^{(0)}$ | $p^{(0)} = \left(\dfrac{1-p}{2-p}, \dfrac{1}{2-p}\right)$ |
| $\eta^{(1)}$ | $p^{(1)} = \left(\dfrac{1}{1+p}, \dfrac{p}{1+p}\right)$ |
| $\eta^{(00)}$ | $p^{(00)} = (0, 1)$ |
| $\eta^{(01)}$ | $p^{(01)} = (1-p, p)$ |
| $\eta^{(10)}$ | $p^{(10)} = (1-p, p)$ |
| $\eta^{(11)}$ | $p^{(11)} = (1, 0)$ |

On the belief 2-simplex, $\eta^{(11)}$, $\eta^{(00)}$, and $\eta^{(01)}$ are the three vertices, while $\eta^{(\emptyset)}$ is the barycentre. The beliefs $\eta^{(0)}$ and $\eta^{(10)}$ lie on the edge joining $\eta^{(11)}$ to $\eta^{(00)}$, and $\eta^{(1)}$ lies on the edge joining $\eta^{(11)}$ to $\eta^{(01)}$.

On the next-token 1-simplex, each point is located by the second coordinate in the table, namely its probability of emitting 1. The relation $p^{(x_{1:n})} = \eta^{(x_{1:n})}A$ maps the two distinct beliefs $\eta^{(01)}$ and $\eta^{(10)}$ to the same point $p^{(01)} = p^{(10)} = (1-p, p)$.

(e) Starting from $\eta^{(10)} = (1-p, p, 0)$, observing a 0 can only come from phase 0 and moves the process to phase 1, which cannot emit a second 0. Hence

$$\Pr(X_{3:4} = 00 \mid X_{1:2} = 10) = \eta^{(10)}T^{(0)}T^{(0)}\mathbf{1} = 0.$$

Starting from $\eta^{(01)} = (0, 0, 1)$, the first 0 has probability $1-p$ and moves the process to phase 0, which emits the second 0 deterministically. Therefore

$$\Pr(X_{3:4} = 00 \mid X_{1:2} = 01) = \eta^{(01)}T^{(0)}T^{(0)}\mathbf{1} = 1-p.$$

The histories agree about the next symbol but disagree about a two-symbol continuation, so their shared next-token vector is not sufficient for the entire future.

(f) We have

$$\delta = \begin{bmatrix} 1-p & p & -1 \end{bmatrix}.$$

Direct multiplication gives

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

Thus the difference between the beliefs is invisible to the one-step readout but visible to a readout associated with the two-symbol future 00.

**Exercise 4.3 (Nonunifilar generators and large belief-state automata).** For a belief $\eta$ and a symbol $a$ satisfying $\eta T^{(a)}\mathbf{1} > 0$, write

$$F_a(\eta) := \frac{\eta T^{(a)}}{\eta T^{(a)}\mathbf{1}}.$$

Thus the belief-state automaton contains the transition

$$\eta \xrightarrow{a:\eta T^{(a)}\mathbf{1}} F_a(\eta).$$

(a) **The Simple Nonunifilar Source.** Revisit the source from Part 2. With state order $(A, B)$, its symbol matrices and initial belief are

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

(i) For $n \geq 0$, show that

$$\eta^{(10^n)} = \begin{bmatrix} \dfrac{1}{n+1} & \dfrac{n}{n+1} \end{bmatrix}, \qquad \eta^{(0^n)} = \begin{bmatrix} \dfrac{1}{n+2} & \dfrac{n+1}{n+2} \end{bmatrix}.$$

Deduce that $\eta^{(0^n)} = \eta^{(10^{n+1})}$ and observe that $\eta^{(\emptyset)} = \eta^{(10)}$.

(ii) Show that, for every $n \geq 1$,

$$\eta^{(10^n 1)} = \eta^{(1)}.$$

(iii) Draw the first four states of the belief-state automaton and indicate its continuing structure with an ellipsis. Show that the beliefs $\eta^{(10^n)}$ are all distinct, identify their limiting point in the 1-simplex, and explain why the two-state generator produces a countably infinite belief-state automaton.

(b) **Mess3.** Now take hidden-state and symbol sets both equal to $\{0, 1, 2\}$, with

$$\alpha = \frac{1}{5}, \qquad b = \frac{1-\alpha}{2} = \frac{2}{5}, \qquad x = \frac{3}{20}, \qquad y = 1 - 2x = \frac{7}{10},$$

and

$$T^{(0)} = \begin{bmatrix} \alpha y & bx & bx \\ \alpha x & by & bx \\ \alpha x & bx & by \end{bmatrix}, \quad T^{(1)} = \begin{bmatrix} by & \alpha x & bx \\ bx & \alpha y & bx \\ bx & \alpha x & by \end{bmatrix},$$

$$T^{(2)} = \begin{bmatrix} by & bx & \alpha x \\ bx & by & \alpha x \\ bx & bx & \alpha y \end{bmatrix}, \quad \eta^{(\emptyset)} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

(i) Verify that $T = \sum_{a=0}^{2} T^{(a)}$ is row-stochastic and explain why this generator is nonunifilar. Compute $\eta^{(0)}, \eta^{(1)}$, and $\eta^{(2)}$.

(ii) Compute $\eta^{(00)}, \eta^{(01)}$, and $\eta^{(02)}$. Use the cyclic symmetry of the process to sketch the first two generations of the belief-state automaton, including the symbol labels.

(iii) Compute the next-token matrix $A$, with $A_{ia} = \sum_j T_{ij}^{(a)}$, and show that it is invertible. What does this imply about whether distinct beliefs can be merged by the map $p^{(x_{1:n})} = \eta^{(x_{1:n})} A$? Contrast this with Exercise 4.2.

(iv) Using a short program, form

$$\mathcal{B}_d := \left\{\eta^{(x_{1:d})} : x_{1:d} \in \{0,1,2\}^d\right\}, \qquad 0 \le d \le 7.$$

Record the number of distinct beliefs at each depth. Visualise $\mathcal{B}_7$ on the hidden-state 2-simplex, colouring each point by its final symbol. Also visualise the associated next-token vectors on the symbol 2-simplex. Describe the geometric structure that emerges and how the three maps $F_0, F_1, F_2$ generate the outgoing transitions of the belief-state automaton.

(v) Explain why the belief-state automata constructed in this exercise are unifilar even though their generators are not. Compare the ladder-with-resets structure of the Simple Nonunifilar Source with the branching structure of Mess3. Finally, explain why the set of beliefs reached by finite histories is countable even though its closure can have a much richer fractal geometry.

**Solution to Exercise 4.3**

(a) (i) After observing 1, the hidden state is $A$. A path emitting the following $0^n$ either remains in $A$ throughout or moves from $A$ to $B$ after one of the $n$ emissions. These $n+1$ paths have equal weight $2^{-n}$; one ends in $A$ and $n$ end in $B$. Thus

$$\eta^{(1)}(T^{(0)})^n = 2^{-n}\begin{bmatrix}1 & n\end{bmatrix}, \qquad \eta^{(10^n)} = \begin{bmatrix}\frac{1}{n+1} & \frac{n}{n+1}\end{bmatrix}.$$

Starting instead from the uniform initial belief gives

$$\eta^{(\emptyset)}(T^{(0)})^n = 2^{-(n+1)}\begin{bmatrix}1 & n+1\end{bmatrix}, \qquad \eta^{(0^n)} = \begin{bmatrix}\frac{1}{n+2} & \frac{n+1}{n+2}\end{bmatrix}.$$

Hence $\eta^{(0^n)} = \eta^{(10^{n+1})}$, and the case $n=0$ gives $\eta^{(\emptyset)} = \eta^{(10)}$.

(ii) For $n \geq 1$, direct multiplication gives

$$\eta^{(10^n)} T^{(1)} = \begin{bmatrix} \frac{n}{2(n+1)} & 0 \end{bmatrix}.$$

Normalising yields $\eta^{(10^n 1)} = \begin{bmatrix} 1 & 0 \end{bmatrix} = \eta^{(1)}$.

(iii) The automaton begins

$$\eta^{(1)} \xrightarrow{0} \eta^{(10)} \xrightarrow{0} \eta^{(100)} \xrightarrow{0} \eta^{(1000)} \xrightarrow{0} \cdots,$$

and every $\eta^{(10^n)}$ with $n \geq 1$ has a 1-transition back to $\eta^{(1)}$. Since the first coordinate $1/(n+1)$ is different for each $n$, the beliefs are distinct, and

$$\lim_{n \to \infty} \eta^{(10^n)} = \begin{bmatrix} 0 & 1 \end{bmatrix}.$$

This limiting belief is not reached after any finite history. Hence the two hidden generator states induce countably infinitely many reachable beliefs.

(b) (i) Substitution gives $\alpha y = 7/50$, $bx = 3/50$, $\alpha x = 3/100$, and $by = 7/25$. Each entry is positive, and every state has several possible destinations for each emitted symbol, so the generator is nonunifilar. The rows of $T^{(0)} + T^{(1)} + T^{(2)}$ each sum to one. By multiplying the uniform belief and normalising,

$$\eta^{(0)} = \begin{bmatrix} \frac{1}{5} & \frac{2}{5} & \frac{2}{5} \end{bmatrix}, \quad \eta^{(1)} = \begin{bmatrix} \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \end{bmatrix}, \quad \eta^{(2)} = \begin{bmatrix} \frac{2}{5} & \frac{2}{5} & \frac{1}{5} \end{bmatrix}.$$

(ii) Updating $\eta^{(0)}$ by each symbol gives

$$\eta^{(00)} = \begin{bmatrix} \frac{13}{87} & \frac{37}{87} & \frac{37}{87} \end{bmatrix},$$

$$\eta^{(01)} = \begin{bmatrix} \frac{52}{163} & \frac{37}{163} & \frac{74}{163} \end{bmatrix}, \quad \eta^{(02)} = \begin{bmatrix} \frac{52}{163} & \frac{74}{163} & \frac{37}{163} \end{bmatrix}.$$

Cyclically permuting both the state coordinates and symbol labels gives the remaining six depth-two beliefs. The root has three symbol-labelled children, and each of those children again has three symbol-labelled successors.

(iii) Summing over destination states gives

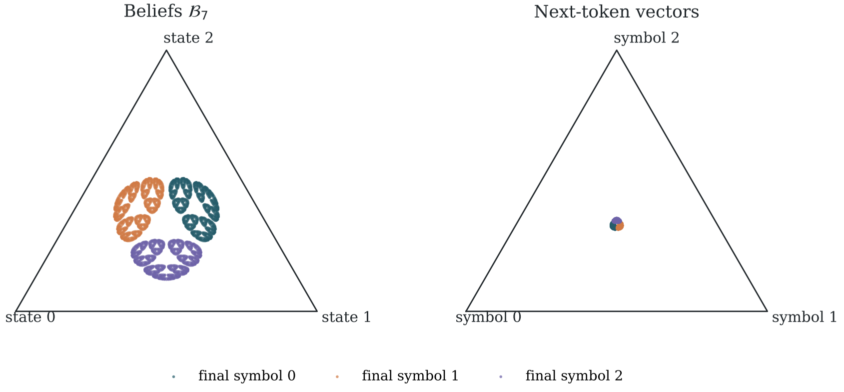
$$A = \begin{bmatrix} \frac{13}{50} & \frac{37}{100} & \frac{37}{100} \\ \frac{37}{100} & \frac{13}{50} & \frac{37}{100} \\ \frac{37}{100} & \frac{37}{100} & \frac{13}{50} \end{bmatrix}.$$

Its eigenvalues are $1, -11/100, -11/100$, so $\det A = 121/10000 \neq 0$. Consequently $\eta A = \eta' A$ implies $\eta = \eta'$: the next-token readout does not merge distinct Mess3 beliefs. In Exercise 4.2 the corresponding matrix has a nontrivial kernel and maps $\eta^{(01)}$ and $\eta^{(10)}$ to the same next-token vector.

(iv) For these parameters, numerical enumeration gives

| $d$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $\lvert\mathcal{B}_d\rvert$ | 1 | 3 | 9 | 27 | 81 | 243 | 729 | 2187 |

with no coincidences at the displayed depths. The plots show three recursively repeated clusters forming a fractal pattern. The next-token plot is an invertible linear image of the belief plot, so it preserves the distinct points and their recursive organisation.




At a point $\eta$, the three outgoing edges are obtained by applying $F_0, F_1, F_2$ and have probabilities given by the three coordinates of $\eta A$. Thus the three geometric copies visible at the next depth are also the three symbol-labelled branches of the belief-state automaton.

(v) A belief and the observed symbol determine a unique updated belief $F_a(\eta)$, so the belief-state presentation is unifilar. The original generator is nonunifilar because its current hidden state and emitted symbol can leave several possible next hidden states. For the Simple Nonunifilar Source the reachable beliefs form a one-dimensional ladder, with each 1-edge resetting to $\eta_0$. For Mess3 every symbol remains possible and the three update maps generate a rapidly branching, self-similar geometry. Finally, finite words form a countable set, so their image under the belief map is countable. Taking the closure adds limiting beliefs and can produce an uncountable fractal set.

**Exercise 4.4 (Generalised hidden Markov models and predictive vectors).** A $d$-dimensional generalised hidden Markov model (GHMM) consists of a finite alphabet $\mathcal{X}$, real $d \times d$ matrices $(T^{(a)})_{a\in\mathcal{X}}$, an initial row vector $\eta^{(\emptyset)}$, and a column vector $\phi$. Writing

$$T := \sum_{a \in \mathcal{X}} T^{(a)},$$

these objects satisfy

$$T\phi = \phi, \qquad \eta^{(\emptyset)}\phi = 1, \qquad \eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} \phi \geq 0$$

for every $x_{1:n} \in \mathcal{X}^n$ and every $n \geq 0$.

(a) **A probability law from matrix products.** Consider

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

(i) Show that these quantities are nonnegative and that

$$\sum_{x_{1:n} \in \mathcal{X}^n} \Pr(X_{1:n} = x_{1:n}) = 1$$

for every $n \geq 0$.

(ii) Show that

$$\sum_{a \in \mathcal{X}} \Pr(X_{1:n+1} = x_{1:n}a) = \Pr(X_{1:n} = x_{1:n}).$$

Explain why this is the consistency condition needed to define a one-sided stochastic process.

(iii) The GHMM conditions do not require $\eta^{(\emptyset)} T = \eta^{(\emptyset)}$. Explain why the process need not therefore be stationary. Show that this additional equality is sufficient for stationarity.

(b) **Predictive vectors.** For any $x_{1:n}$ satisfying $\Pr(X_{1:n} = x_{1:n}) > 0$, define its predictive vector by

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

(i) Show that

$$\eta^{(x_{1:n})}\phi = 1$$

and, for every continuation $y_{1:m}$,

$$\Pr(X_{n+1:n+m} = y_{1:m} \mid X_{1:n} = x_{1:n}) = \eta^{(x_{1:n})} T^{(y_1)} \cdots T^{(y_m)} \phi.$$

(ii) Derive the recursive update and identify its normalising factor:

$$\eta^{(x_{1:n}a)} = \frac{\eta^{(x_{1:n})} T^{(a)}}{\eta^{(x_{1:n})} T^{(a)} \phi}, \quad \Pr(X_{n+1} = a \mid X_{1:n} = x_{1:n}) = \eta^{(x_{1:n})} T^{(a)} \phi.$$

(iii) Conceptually contrast an HMM belief state with a GHMM predictive vector. Compare the meaning of their coordinates, their positivity and normalisation properties, and how they encode predictions of future observations.

(c) **Predictive vectors need not be probability vectors.** Consider the two-state HMM

$$M^{(0)} = \begin{bmatrix} \frac{4}{5} & 0 \\ 0 & \frac{1}{5} \end{bmatrix}, \qquad M^{(1)} = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{4}{5} \end{bmatrix}, \qquad \alpha = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Thus a hidden state is selected initially and thereafter emits a biased coin without changing state. Let

$$S = \begin{bmatrix} 2 & -1 \\ 0 & 1 \end{bmatrix}, \qquad T^{(a)} := S^{-1} M^{(a)} S, \qquad \eta^{(\emptyset)} := \alpha S, \qquad \phi := S^{-1}\mathbf{1}.$$

(i) Compute $T^{(0)}, T^{(1)}, \eta^{(\emptyset)}$, and $\phi$. Show that, for every $x_{1:n}$,

$$\eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} \phi = \alpha M^{(x_1)} \cdots M^{(x_n)} \mathbf{1}.$$

Deduce that this change of coordinates gives a valid GHMM for the same process.

(ii) Compute $\eta^{(0)}$ and $\eta^{(1)}$. Verify directly from $\eta^{(0)}$ that

$$\Pr(X_2 = 0 \mid X_1 = 0) = \frac{17}{25}, \qquad \Pr(X_2 = 1 \mid X_1 = 0) = \frac{8}{25}.$$

Explain why the entries of a GHMM predictive vector need not themselves be probabilities.

(d) **The probability matrix and minimal dimension.** For finite sets $\mathcal{U}, \mathcal{V} \subseteq \mathcal{X}^*$, define the probability matrix by

$$\mathsf{P}_{\mathcal{U},\mathcal{V}}(u, v) := \Pr(X_{1:|u|+|v|} = uv), \qquad u \in \mathcal{U}, \quad v \in \mathcal{V}.$$

(i) Use the GHMM matrix product to factor $\mathsf{P}_{\mathcal{U},\mathcal{V}}$ as a matrix of prefix row vectors times a matrix of continuation column vectors. Deduce that

$$\operatorname{rank} \mathsf{P}_{\mathcal{U},\mathcal{V}} \leq d$$

for every $d$-dimensional GHMM realisation.

(ii) For the process in part (c), take $\mathcal{U} = \mathcal{V} = \{\emptyset, 0, 1\}$. Compute $\mathsf{P}_{\mathcal{U},\mathcal{V}}$ and use an SVD to find its singular values. Deduce that the two-dimensional GHMM is minimal-dimensional.

**Solution to Exercise 4.4**

(a) (i) Nonnegativity is one of the GHMM assumptions. Expanding the product of the sum of the symbol matrices gives

$$\sum_{x_{1:n} \in \mathcal{X}^n} T^{(x_1)} \cdots T^{(x_n)} = T^n.$$

Hence

$$\sum_{x_{1:n} \in \mathcal{X}^n} \Pr(X_{1:n} = x_{1:n}) = \eta^{(\emptyset)} T^n \phi = \eta^{(\emptyset)} \phi = 1,$$

where $T^n \phi = \phi$ follows from $T\phi = \phi$. For $n = 0$ this also says that the empty history has probability one.

(ii) Summing over the final symbol gives

$$\begin{aligned}
\sum_{a \in \mathcal{X}} \Pr(X_{1:n+1} = x_{1:n}a) &= \eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} \left( \sum_{a \in \mathcal{X}} T^{(a)} \right) \phi \\
&= \eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} T \phi \\
&= \Pr(X_{1:n} = x_{1:n}).
\end{aligned}$$

Thus the length-$n$ distribution is the marginal of the length-$(n+1)$ distribution over its final coordinate, as required for a process indexed by $1, 2, \ldots$.

(iii) The displayed GHMM conditions relate distributions at successive lengths but do not say that a block has the same distribution after shifting its time indices. If $\eta^{(\emptyset)} T = \eta^{(\emptyset)}$, then

$$\begin{aligned}
\sum_{a \in \mathcal{X}} \Pr(X_{1:n+1} = a x_{1:n}) &= \eta^{(\emptyset)} T T^{(x_1)} \cdots T^{(x_n)} \phi \\
&= \eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} \phi \\
&= \Pr(X_{1:n} = x_{1:n}).
\end{aligned}$$

The left side is $\Pr(X_{2:n+1} = x_{1:n})$, so the block distributions are invariant under a one-step shift. Iteration gives stationarity.

(b) (i) The denominator in the definition is $\Pr(X_{1:n} = x_{1:n})$, and so

$$\eta^{(x_{1:n})} \phi = 1.$$

By the definition of conditional probability,

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

(ii) Taking $m = 1$ in part (i) gives

$$\Pr(X_{n+1} = a \mid X_{1:n} = x_{1:n}) = \eta^{(x_{1:n})} T^{(a)} \phi, \quad \eta^{(x_{1:n}a)} = \frac{\eta^{(x_{1:n})} T^{(a)}}{\eta^{(x_{1:n})} T^{(a)} \phi}.$$

The first quantity is exactly the normalising factor in the update shown by the second equality.

(iii) An HMM belief state has coordinates $\Pr(S_n = i \mid X_{1:n} = x_{1:n})$, so they are nonnegative, sum to one, and refer to mutually exclusive hidden states. Predictive vectors are coordinates in a linear realisation: they need only satisfy $\eta^{(x_{1:n})}\phi = 1$ and may have negative entries. Both summarise the history sufficiently to compute future-word probabilities through symbol-matrix products.

(c) (i) Since

$$S^{-1} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{bmatrix},$$

direct calculation gives

$$T^{(0)} = \begin{bmatrix} \frac{4}{5} & -\frac{3}{10} \\ 0 & \frac{1}{5} \end{bmatrix}, \qquad T^{(1)} = \begin{bmatrix} \frac{1}{5} & \frac{3}{10} \\ 0 & \frac{4}{5} \end{bmatrix},$$

and

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

In particular $T = T^{(0)} + T^{(1)} = I$, so $T\phi = \phi$, and $\eta^{(\emptyset)}\phi = 1$. In a product, adjacent factors $SS^{-1}$ cancel:

$$\begin{aligned} \eta^{(\emptyset)} T^{(x_1)} \cdots T^{(x_n)} \phi &= \alpha S (S^{-1} M^{(x_1)} S) \cdots (S^{-1} M^{(x_n)} S) S^{-1} \mathbf{1} \\ &= \alpha M^{(x_1)} \cdots M^{(x_n)} \mathbf{1}. \end{aligned}$$

The right side is an HMM sequence probability and is therefore nonnegative. All GHMM conditions are satisfied, and both realisations define the same process.

(ii) Both first-symbol probabilities are $1/2$. Normalising gives

$$\eta^{(0)} = \begin{bmatrix} \frac{8}{5} & -\frac{3}{5} \end{bmatrix}, \qquad \eta^{(1)} = \begin{bmatrix} \frac{2}{5} & \frac{3}{5} \end{bmatrix}.$$

Moreover,

$$T^{(0)}\phi = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{5} \end{bmatrix}, \qquad T^{(1)}\phi = \begin{bmatrix} \frac{1}{2} \\ \frac{4}{5} \end{bmatrix}.$$

Hence

$$\eta^{(0)} T^{(0)} \phi = \frac{17}{25}, \qquad \eta^{(0)} T^{(1)} \phi = \frac{8}{25}.$$

Although the second coordinate of $\eta^{(0)}$ is negative, all continuation probabilities obtained from it are valid. The predictive vector consists of coordinates in a chosen linear representation; unlike an HMM belief, its entries do not have to describe mutually exclusive hidden events.

(d) (i) For $u = u_1 \cdots u_r$ and $v = v_1 \cdots v_s$,

$$\mathsf{P}_{\mathcal{U},\mathcal{V}}(u, v) = \left(\eta^{(\emptyset)} T^{(u_1)} \cdots T^{(u_r)}\right)\left(T^{(v_1)} \cdots T^{(v_s)} \phi\right).$$

Collecting the first factors as rows and the second factors as columns gives the requested factorisation through $\mathbb{R}^d$. The rank of the product is therefore at most $d$.

(ii) The required probabilities are

$$\begin{aligned} \Pr(X_1 = 0) = \Pr(X_1 = 1) &= \frac{1}{2}, \\ \Pr(X_{1:2} = 00) = \Pr(X_{1:2} = 11) &= \frac{17}{50}, \\ \Pr(X_{1:2} = 01) = \Pr(X_{1:2} = 10) &= \frac{4}{25}. \end{aligned}$$

Thus, in the order $(\emptyset, 0, 1)$,

$$\mathsf{P}_{\mathcal{U},\mathcal{V}} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{17}{50} & \frac{4}{25} \\ \frac{1}{2} & \frac{4}{25} & \frac{17}{50} \end{bmatrix}, \qquad \sigma(\mathsf{P}_{\mathcal{U},\mathcal{V}}) = \left(\frac{3}{2}, \frac{9}{50}, 0\right).$$

It therefore has rank two. Every GHMM realisation must have dimension at least two, while part (c) supplies a two-dimensional realisation. It is minimal-dimensional.

## 5 Identifying belief geometry in transformers

So far, we have developed the conceptual framework of computational mechanics. We now ask whether neural networks trained for next-token prediction—including transformers and recurrent architectures—learn the predictive structures identified by that framework. You will read selected sections of two papers, then discuss their evidence and limitations in small groups.

**Session at a glance**

1. **Readings (75 minutes).** Begin with the main reading and use the route below. Continue to the extension if time permits.

2. **Discussion (30 minutes).** Form groups of three or four. Compare what

you found convincing, confusing, or incomplete; use the prompts on page 2 if useful.

## 5.1 Readings

Prioritise the main reading. If you can answer its evaluation question, you have probably understood one of its central results. If time is short, skim the extension overview before deciding whether to continue.

### 5.1.1 [Main] Transformers represent belief state geometry in their residual stream

**Reading route**

- **Skim.** Sections 1, 2.1, and 2.2.
- **Read.** Sections 2.3 and 3–5.
- **Evaluation.** Explain Figure 6D. What do the comparison and shuffle control establish?

**Overview.** This paper presents evidence that transformers trained on HMM-generated sequences encode belief-state geometry in their residual streams. The claim is tested on the Mess3 and Random–Random–XOR (RRXOR) processes. For each process, the authors fit an affine map from residual activations to the exact HMM belief state and evaluate it using mean-squared error (MSE). The error decreases over training checkpoints. A shuffle control tests whether low MSE could arise merely because a high-dimensional activation space is being projected into a low-dimensional belief space.

### 5.1.2 [Extension] Neural networks leverage nominally quantum and post-quantum representations

**Reading route**

- **Skim.** Sections 1–3.
- **Read.** Sections 4–8.
- **Evaluation.** Explain Figure 3B. How are the light-blue and light-orange plots related to the corresponding dark-blue and dark-orange plots?

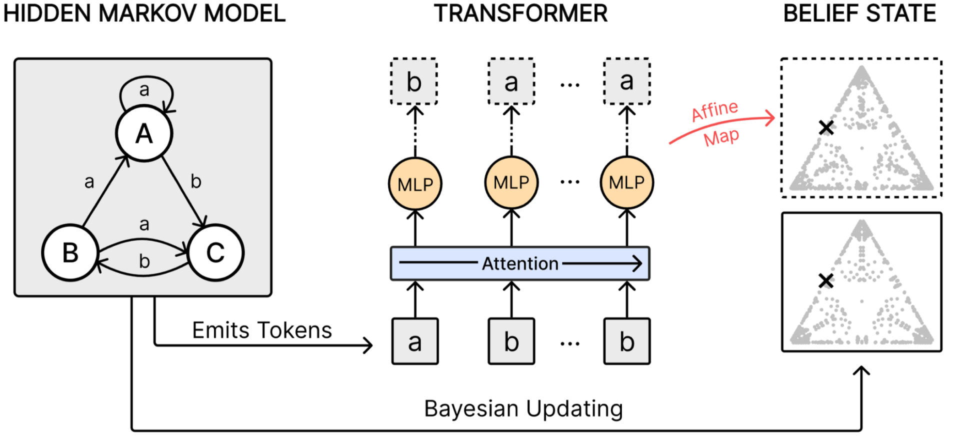
**Overview.** This paper extends the question from HMMs to generalised hidden Markov models (GHMMs). As in Part 3, a GHMM retains the matrix-product expression for output probabilities while relaxing the entrywise probabilistic constraints on its internal matrices. This larger model class can realise some processes in finite dimension even when a corresponding minimal HMM requires exponentially many more hidden states.

Using techniques related to those in the main reading, the authors study transformers, LSTMs, RNNs, and GRUs trained on processes with distinct GHMM and HMM realisations. Their results indicate that the networks learn belief-state representations associated with the lower-dimensional GHMM realisation. The finding suggests that neural networks can select among representational model classes in ways that capture dimensionality savings.

### 5.2 Discussion

Here are some prompts to consider:

- What is the most convincing evidence that transformers represent belief-state geometry? What is the least convincing?
- What additional experiment would most strengthen or weaken the claim?
- Which alternative hypotheses could explain the reported results?
- Can you explain each element of the diagram below—including the generative HMM, Bayesian updating, transformer activations, fitted affine map, and belief-state geometry?




## 6 Python exercises: HMM from scratch

The Python exercises are ARENA-style notebooks in two parts. Each part has an **exercises** notebook (with `# YOUR CODE HERE` stubs, inline tests, and collapsible solutions) and a fully worked **solutions** notebook.

- **Part 1 — Sequence probabilities & next-token distributions** (exercises) — Open in Colab

- Part 1 (solutions) — Open in Colab
- **Part 2 — Belief states** (exercises) — Open in Colab
- Part 2 (solutions) — Open in Colab

The Colab notebooks are self-contained: their setup cells write the helper modules into the session, so there is nothing else to install or download.

# 7 Learn more

- Beyond toy architectures – in-context learning
  - LLMs have an amazing ability to adapt internal representations in-context – [Read: ICLR: In-Context Learning of Representations]
  - LLMs can predict HMM data in-context – [Read: Pre-trained Large Language Models Learn Hidden Markov Models In-context]
  - This suggests that internal representations of LLMs predicting GHMM data in-context, may resemble belief states.
- Processes consisting of GHMMs with richer structure [Read: Transformers learn factored representations]
- Review article of computational mechanics [Read: Between order and chaos]
- Simplex is a non-profit research organisation working on applying computational mechanics to AI safety.