ILIAD Intensive

Computational Mechanics

2. Predicting the future

Xavier Poncini (Simplex)

Autumn 2026

Outline

1. Sufficient statistics for prediction
2. Causal states: minimal predictive summaries
3. Causal-state dynamics
4. Next-token prediction in RNNs and Transformers

Predicting the future from the past

Let $(X_t)_{t\geq 1}$ be a discrete-time stochastic process on a finite alphabet $\mathcal{X}$.

For consecutive times i to j, we introduce the notation:

$$X_{i:j} = (X_i, X_{i+1}, \ldots, X_j), \qquad x_{i:j} = (x_i, x_{i+1}, \ldots, x_j) \in \mathcal{X}^{j-i+1}.$$

Consider predicting the **future** $X_{m+1:m+n}$ from **past** observations $X_{1:m} = x_{1:m}$:

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

How should we do this **efficiently**?

Condensing the past

The **possible histories** are the words that can occur:

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

A **statistic** maps allowable sequences to summary values:

$$R : \mathcal{X}_+^* \to \mathcal{R}.$$

R is **sufficient for prediction** if histories assigned the same summary induce the same future distribution:

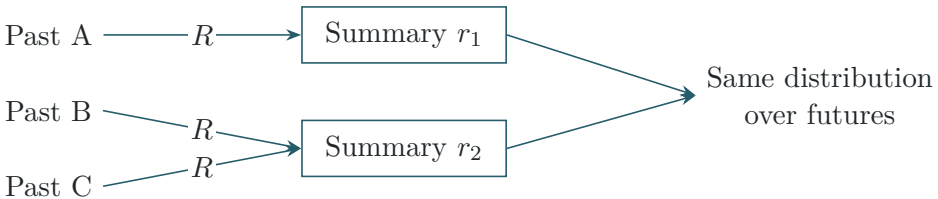
$$R(x) = R(x') \implies$$

$$\Pr\Big(X_{|x|+1:|x|+|y|} = y \mid X_{1:|x|} = x\Big) = \Pr\Big(X_{|x'|+1:|x'|+|y|} = y \mid X_{1:|x'|} = x'\Big).$$

This must hold for every $x, x' \in \mathcal{X}_+^*$ and every $y \in \mathcal{X}^*$.

Sufficient statistics

A sufficient statistic can still retain unnecessary distinctions between pasts.



Two summary values give identical predictions: **sufficient, but not minimal**.

Shalizi, Lecture I (1998), §3.2; Shalizi & Crutchfield (2001), §V.A–B, cond-mat/9907176.

Minimal sufficient statistics

Merge summary values with identical distributions over futures.

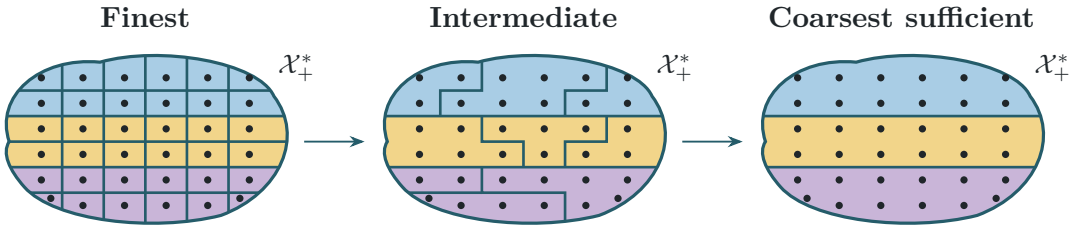
r_1 and r_2 merge to r under the new statistic $R_{\min}$: predictions are unchanged.

A **minimal sufficient statistic** gives the **coarsest grouping of pasts** that remains sufficient. No two remaining summary values can be merged without changing predictions.

Shalizi, Lecture I (1998), §3.2; Shalizi & Crutchfield (2001), §V.A–B, cond-mat/9907176.

Sufficient statistics as partitions

Sufficient statistics group input sequences with the same future distribution.



Minimal sufficient statistics are **memory efficient**.

Schematic: dots represent example sequences; matching colours indicate identical predictions.

Causal states

For each $x \in \mathcal{X}_+^*$, define the map from histories to classes of histories by:

$$\epsilon(x) \equiv \left\{ x' \in \mathcal{X}_+^* \;\middle|\; \begin{array}{c} \text{for every } y \in \mathcal{X}^*, \\ \Pr\Big(X_{|x|+1:|x|+|y|} = y \mid X_{1:|x|} = x\Big) = \Pr\Big(X_{|x'|+1:|x'|+|y|} = y \mid X_{1:|x'|} = x'\Big) \end{array} \right\}.$$

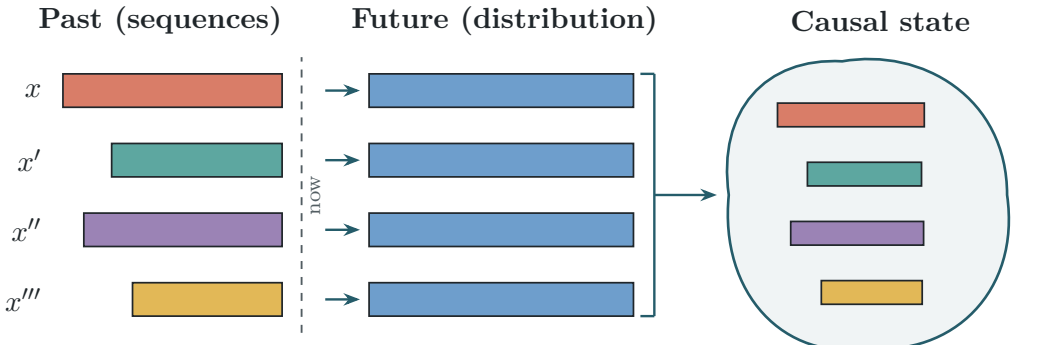
The values of ϵ are the (finite-history) **causal states**.

"The causal states record every distinction that makes a difference."

Theorem. ϵ is a minimal sufficient statistic for prediction.
Proof. We will prove this result in the exercises.

Upper, *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models* (1997), §§2.4–2.5; Shalizi, Lecture I (1998), §3.1.

Past sequences belong to the same causal state when they induce the **same future distribution**.



Same future law $\iff$ same causal state.

Example 1: (IID process)

Emits sequences over $\mathcal{X} = \{0, 1\}$ by sampling each symbol independently from the same distribution, with both symbols having positive probability.

There is one causal state:

$$\epsilon(x) = \{(),\ 0,\ 1,\ 00,\ 01,\ 10,\ 11, \ldots\} = \mathcal{X}_+^* = \mathcal{X}^*, \qquad x \in \mathcal{X}^*.$$

Example 2: (Golden Mean process)

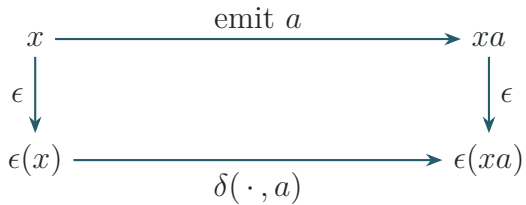
Emits sequences over $\mathcal{X} = \{0, 1\}$ in which consecutive 0s are forbidden and, after 1, either symbol can occur. No earlier symbols affect emission probabilities.

For nonempty histories, there are two causal states:

$$\epsilon(0) = \{0,\ 10,\ 110,\ 010, \ldots\} = \left\{x \in \mathcal{X}_+^* \mid x \text{ ends in } 0\right\},$$
$$\epsilon(1) = \{1,\ 01,\ 11,\ 101, \ldots\} = \left\{x \in \mathcal{X}_+^* \mid x \text{ ends in } 1\right\}.$$

Updating causal states

Now consider updating the causal state as the process emits a new token $a \in \mathcal{X}$:



Process

Predictive state

Proposition. The assignment

$$\delta(\epsilon(x), a) := \epsilon(xa), \qquad xa \in \mathcal{X}_+^*,$$

is well-defined: it depends only on the causal state $\epsilon(x)$ and the emitted token a.

Corollary. Causal-state dynamics form a **partial deterministic automaton**.

Upper (1997), §2.4; Shalizi & Crutchfield (2001), §IV.C, Lemma 5.

Proof of the proposition

Proof. We must show that the value assigned by $\delta(\,\cdot\,, a)$ is independent of the representative chosen from a causal state.

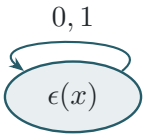
Let $x, x' \in \mathcal{X}_+^*$ satisfy $\epsilon(x) = \epsilon(x')$, and suppose $xa \in \mathcal{X}_+^*$. Predictive equivalence gives $\Pr(a \mid x) = \Pr(a \mid x') > 0$, so $x'a$ is also possible. For every $y \in \mathcal{X}^*$,

$$\Pr(y \mid xa) = \frac{\Pr(ay \mid x)}{\Pr(a \mid x)} = \frac{\Pr(ay \mid x')}{\Pr(a \mid x')} = \Pr(y \mid x'a).$$

Thus xa and $x'a$ induce the same future distribution, so $\epsilon(xa) = \epsilon(x'a)$. Therefore $\delta(\,\cdot\,, a)$ assigns the same image regardless of the representative chosen from the causal state. □

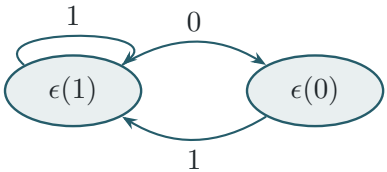
Example 1: (IID process)

Emits sequences over $\mathcal{X} = \{0, 1\}$ by sampling each symbol independently from the same distribution, with both symbols having positive probability.



Example 2: (Golden Mean process)

Emits sequences over $\mathcal{X} = \{0, 1\}$ in which consecutive 0s are forbidden and, after 1, either symbol can occur. No earlier symbols affect emission probabilities.



Taking stock

For predicting the **entire future**:

- Causal states are the **most efficient summary of the past**.
- Given a new observation, the causal state is **easy to update**.

> But modern autoregressive neural networks do not predict the entire future directly. They predict **only the next-token distribution**.

Recurrent next-token prediction

Given hidden state h_{n-1} and observation x_n, a model with parameters θ recurrently updates its hidden state and produces prediction y_n:

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

Proposition. Fix the initial hidden state h_0. Suppose that, for every $n \geq 1$,

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

Then the recurrent state (h_{n-1}, x_n) is a sufficient statistic for the entire future.

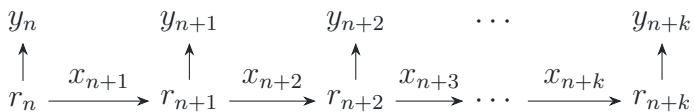
Exact recurrent **next-token prediction**
$\Longrightarrow$ a sufficient statistic for the **entire future!**

Shalizi (2001), Theorem 6; Shalizi & Shalizi (2004), §2.

Why is next-token prediction enough?

We leave the proof as an exercise and provide the intuition here.

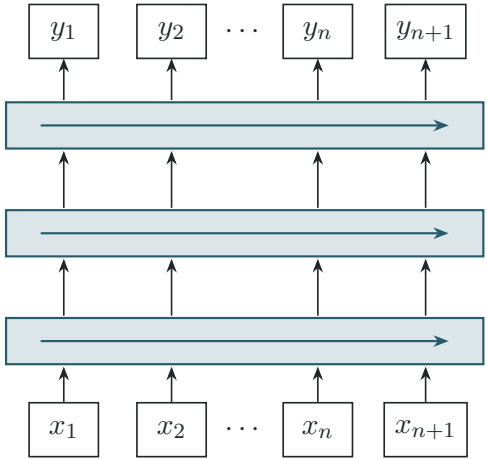
Let $r_n := (h_{n-1}, x_n)$ and $y_n := \Pr(X_{n+1} \mid X_{1:n} = x_{1:n})$. The recurrent update determines the entire future distribution via next-token probabilities:



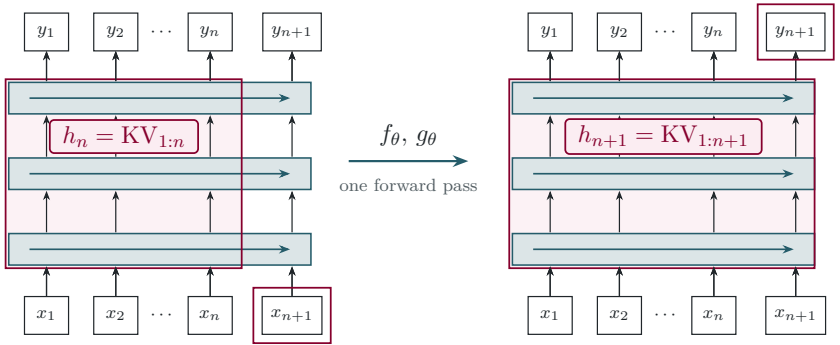
Information in the past $x_{1:n}$ inducing a future distinction must be stored in r_n.

A. Shai (2026), private communication.

What about transformers?



What about transformers?



Transformer with **exact next-token prediction**

$\Longrightarrow$ (h_n, x_{n+1}) is sufficient for **the entire future!**

Dai et al. (2019); Katharopoulos et al. (2020, linear attention); A. Kolchinsky (2026), private communication.

Do neural networks learn causal states?

Recurrent neural networks and Transformers that achieve exact next-token prediction develop **sufficient statistics** for the entire future!

Why might these sufficient statistics also be minimal?

Simplicity bias: Capacity constraints and regularisation favour simpler representations that discard distinctions irrelevant to prediction.

How might a neural network encode the structure of a causal state?

In the following lecture, we will turn to this question.

Summary

For predicting the entire future:

- Causal states are the **most efficient summary of the past**.
- Given a new observation, the causal state is **easy to update**.

For next-token prediction:

- Recurrent neural networks and Transformers that achieve exact next-token prediction develop **sufficient statistics** for the entire future.
- Simplicity bias suggests that these learned statistics may approach **minimal sufficient statistics**.