BEYOND LAPLACE: HOW MAMBA IN-CONTEXT LEARNS CONSTRAINED SWITCHING MARKOV CHAINS — ESTIMATION GAPS CLOSE, STATISTICAL LIMITS DO NOT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work showed that Mamba exactly represents, and learns in training, the Laplacian smoothing estimator for unconstrained Markov chains. We study switching Markov chains whose transition kernels share one stationary distribution and whose switches are hidden from the predictor. This setting poses two distinct problems, estimating the constrained kernel and identifying the kernel currently in effect. For estimation, the stationarity constraint couples the kernel rows in the posterior, so Laplacian smoothing is strictly suboptimal for any number of states. We show that the difference between the constrained Bayes predictor and Laplacian smoothing lies in the readout rather than in the state. Readouts that use only the current row's counts cannot reach this predictor, whereas a log-linear softmax readout on the same count state represents it exactly for two states, with a width that grows linearly in the sequence length and a lower bound of the same order. Probes show that trained Mamba stores the transition counts linearly in its hidden state and combines counts across rows at the readout, closing most of the estimation gap. For identification, even the optimal filter with infinite memory incurs a strictly positive residual loss for any number of states and any order, and a Wald-type lower bound shows that switches cannot be detected instantly. Trained Mamba and Transformer models approach this statistical limit but do not exceed it. Provable unpredictability against well-trained sequence models must therefore rest on the statistical limit created by hidden switching rather than on the difficulty of estimation. Code is available at `https://anonymous.4open.science/r/BeyondLaplace-6CC7`.

1 INTRODUCTION

Random Markov chains provide a controlled setting for studying in-context learning in sequence models (Edelman et al., 2024; Nichani et al., 2024; Rajaraman et al., 2024a). If the rows of the transition kernel independently follow Dirichlet priors, the Bayes-optimal predictor is the Laplacian smoothing estimator. In this setting, Transformers form statistical induction heads and approach the Bayes optimum (Edelman et al., 2024), and they need at least two layers to do so (Sanford et al., 2024; Ekbote et al., 2025). Empirical conclusions on the in-context learning of selective state space models such as Mamba (Gu & Dao, 2023; Dao & Gu, 2024) are mixed (Grazzi et al., 2024; Park et al., 2024; Akyürek et al., 2024). Bondaschi et al. (2026) proved that a single-layer Mamba can exactly represent the Laplacian smoothing estimator, and that trained models converge to this estimator.

These results rest on two premises. First, the rows of the transition kernel are mutually independent under the prior. Second, the generating process is static, or its switches are visible to the model. Systems that must maintain the long-run visitation frequency of each state (Alvarenga et al., 2019; Wang et al., 2026) violate both premises (Appendix E). Their transition kernels are constrained to have a given stationary distribution, and they may switch invisibly among several transition kernels that satisfy this constraint. In this case Laplacian smoothing is no longer Bayes optimal, and the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

model must also infer the transition kernel currently in effect. This paper studies what Mamba learns in this situation, and what no predictor can learn.

We study switching Markov chains with a stationarity constraint, which contain two problems, estimating the constrained transition kernel and identifying the current transition kernel (Figure 1(a)(b)). The posterior predictive mean under the stationarity-constrained prior is called the constrained Bayes predictor. The gap between Laplacian smoothing and this predictor is a capability gap, that is, a gap that a sufficiently strong predictor can eliminate. The oracle predictor is the predictor that knows the current transition kernel. The loss difference between the optimal filter and the oracle predictor is called the residual loss, and it is a statistical limit, that is, a lower bound on the loss that holds for every predictor. The main contributions of this paper are as follows.

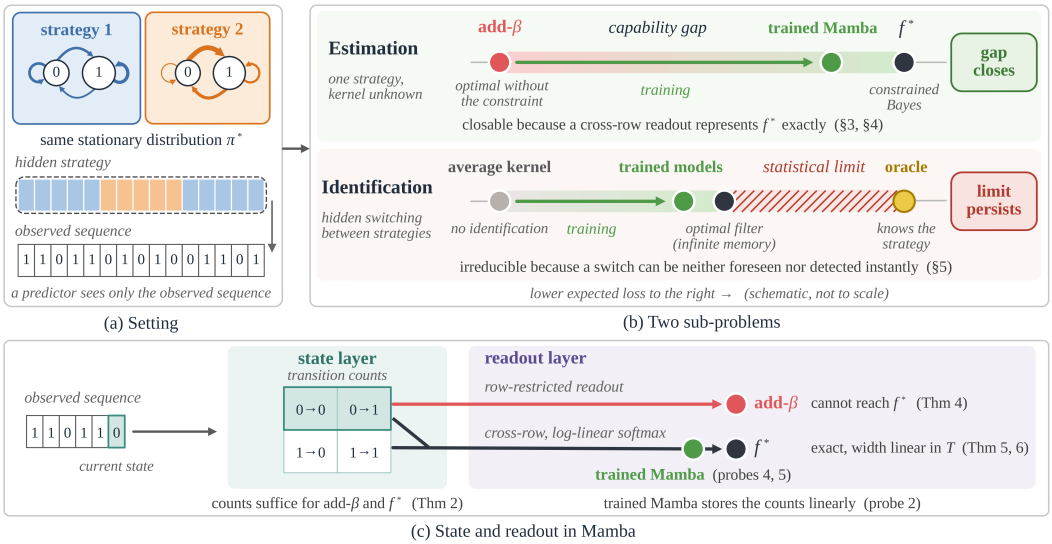


Figure 1: Framework of this paper. (a) Setting. (b) The two subproblems of estimation and identification, where the horizontal axis is a schematic of the expected loss. (c) The state layer of Mamba and two readouts.

- **Strict suboptimality of Laplacian smoothing under the constraint (Section 3, Theorems 1 to 3).** The stationarity constraint couples the rows of the transition kernel in the posterior, so Laplacian smoothing is strictly suboptimal for any number of states. We also give the closed form for two states, the asymptotic order of the gap, and the structural component under switching.
- **Impossibility and constructive representation at the readout (Section 4, Theorems 4 to 6, Figure 1(c)).** Readouts whose input is only the counts of the current row cannot reach the constrained Bayes predictor, whereas a log-linear softmax readout represents it exactly for two states, with a required width of the same order as the sequence length.
- **Trained Mamba learns a cross-row readout (Sections 3.4 and 4.1, Appendix C).** After training on constrained data, all four architectures outperform Laplacian smoothing, and Gated M-2 and full Mamba close most of the estimation gap. Probes show that Mamba stores the transition counts linearly and uses the counts of all rows at the readout.
- **Statistical limit under hidden switching (Section 5, Theorem 7 and Proposition 2).** Even with infinite memory, the residual loss of the optimal filter is strictly positive for any number of states and any order, and switches cannot be detected instantly. All trained models approach the optimal filter but do not exceed it.

Related work. Prior work has studied statistical induction heads, training dynamics, the required depth, the effect of tokenization, and the connection between self-attention and Markov models for Transformers on Markov chains (Edelman et al., 2024; Nichani et al., 2024; Rajaraman et al., 2024a; Sanford et al., 2024; Ekbote et al., 2025; Makkuva et al., 2024; 2025; Rajaraman et al.,

2024b; Ildiz et al., 2024). This line of work has been extended to variable-order Markov chains and to the case in which the training and test transition matrices differ (Zhou et al., 2026; Lepage et al., 2025). In these works the transition kernel is either fixed or has independently sampled rows, whereas this paper studies the case in which the rows are coupled through a shared stationary distribution. Induction heads are regarded as the core circuit of in-context learning in Transformers (Olsson et al., 2022; Reddy, 2024). Comparisons between architectures reach mixed conclusions (Grazzi et al., 2024; Park et al., 2024; Akyürek et al., 2024; Halloran et al., 2025), and architectures with similar performance may still differ in their internal mechanisms (Wang et al., 2025). The in-context learning of linear recurrent models and the feature learning and expressive power of Mamba have also been analyzed theoretically (Tian et al., 2026; Oh et al., 2025; Muca Cirone et al., 2024). The behavior of sequence models on hidden Markov data and on latent mixture data has been studied (Shai et al., 2024; Hu et al., 2024; D'Angelo & Flammarion, 2026), and detection delay is constrained by sequential detection theory (Wald, 1945; Page, 1954; Lorden, 1971; Lai, 1998). The complete related work is given in Appendix G.

2 Problem setup

2.1 Constrained switching Markov chains

State process. The observable state process takes values in $\mathcal{X} = \{0, \dots, n-1\}$. A transition kernel P is an $n \times n$ row-stochastic matrix with $P_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i)$. The target stationary distribution $\pi^* \in \Delta^{n-1}$ represents the long-run frequency with which each state is visited, and the constraint set is $\mathcal{C}(\pi^*) = \{P \text{ row-stochastic} : \pi^* P = \pi^*\}$. The **strategy set** $\mathcal{P} = \{P^{(1)}, \dots, P^{(K)}\} \subset \mathcal{C}(\pi^*)$ consists of the transition kernels available to the data-generating process. Its elements are called strategies, and K is the number of strategies.

Strategy process. The unobservable index of the strategy in effect, $Z_t \in \{1, \dots, K\}$, evolves according to the switching matrix A, with $\mathbb{P}(Z_{t+1} = \ell' \mid Z_t = \ell) = A_{\ell\ell'}$. Let w denote the stationary distribution of A and p_{switch} the total switching probability of A, which for $K = 2$ is the off-diagonal entry $A_{12} = A_{21}$.

Joint dynamics. Initially $X_1 \sim \pi^*$ and $Z_1 \sim w$, and the two are independent. Thereafter $X_{t+1} \sim P^{(Z_{t+1})}(\cdot \mid X_t)$.

We define the strategy-averaged kernel $\bar{P}^w = \sum_\ell w_\ell P^{(\ell)}$, the within-strategy entropy $H_{\text{within}} = \sum_{\ell=1}^K w_\ell H_{\pi^*}(P^{(\ell)})$ and the between-strategy entropy $H_{\text{between}} = H_{\pi^*}(\bar{P}^w) - H_{\text{within}} = I(X_{t+1}; Z_{t+1} \mid X_t)$, where $H_{\pi^*}(P) = \sum_i \pi_i^* H(P_{i\cdot})$. The quantity H_{within} is the loss of the oracle predictor, and H_{between} measures the difference between strategies and is zero when all strategies are identical. Each theorem invokes the parts it needs of the following assumptions.

- (A1) π^* lies in the interior of the probability simplex, that is, $\pi_i^* > 0$ for all $i \in \mathcal{X}$.
- (A2) The prior pseudo-count satisfies $\beta > 0$, and the prior has a strictly positive density on the relative interior of $\mathcal{C}(\pi^*)$.
- (A3) Every transition kernel in $\mathcal{P}$ is irreducible, all entries of A are strictly positive, and $p_{\text{switch}} \in (0, 1)$.

2.2 Prediction model

A predictor outputs a probability distribution over the next state from the observed state sequence. It knows $\mathcal{X}$, π^* and $\mathcal{C}(\pi^*)$, and it observes neither Z_t nor the switching times. In the single-strategy setting (Sections 3 and 4) the transition kernel P is unknown. In the switching setting (Section 5) $\mathcal{P}$ and A are known, so the residual loss comes entirely from identification.

Let $\mathcal{L}_t(f) = \mathbb{E}[-\log f_t(x_{t+1} \mid x_1^t)]$ denote the expected log loss of predictor f at step t, where the expectation is taken jointly over the random kernel and the state sequence. The per-step loss $\mathcal{L}(f) = \frac{1}{T}\sum_{t=1}^T \mathcal{L}_t(f)$ is the average of the losses over the first T steps. The corresponding optimal

losses are $\mathcal{L}_t^* = \min_f \mathcal{L}_t(f)$ and $\mathcal{L}^* = \min_f \mathcal{L}(f) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_t^*$, and $\mathrm{excess}_t(f) = \mathcal{L}_t(f) - \mathcal{L}_t^*$ is the excess risk at step t.

The degree ρ to which a predictor identifies the strategy currently in effect is defined by $\mathcal{L}(f) = H_{\text{within}} + (1-\rho)\, H_{\text{between}}$. Here $\rho = 1$ corresponds to complete identification, and $\rho = 0$ is equivalent to using the strategy-averaged kernel $\bar{P}^w$.

Laplacian smoothing. Laplacian smoothing is denoted add-β. It uses the count n_{ij} of $i \to j$ transitions in the sequence and predicts $f_\beta(x_{t+1} = j \mid x_1^t) = (n_{ij} + \beta)/(n_{i\cdot} + n\beta)$, where $i = x_t$ is the current state and $n_{i\cdot} = \sum_j n_{ij}$. It is the Bayes-optimal and asymptotically minimax-optimal predictor under a product Dirichlet prior with independent rows (Xie & Barron, 1997; Hao et al., 2018).

Architectures tested and evaluation metric. The structures and training settings of the four tested architectures, the MambaZero variant, Gated M-2, full Mamba and the Transformer, are given in Appendix D. The closed fraction of the estimation gap is defined as $100\% \times (\mathcal{L}(f_\beta) - \mathcal{L}_{\text{model}})/(\mathcal{L}(f_\beta) - \mathcal{L}(f^*))$, where 0 and 100% correspond to the losses of add-β and f^*, respectively.

3 Estimation gap under the stationarity constraint

3.1 Constraint set and prior

Let $n = 2$, $\pi^* = (\pi_0, \pi_1)$, and let the transition kernel be $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$, where $p = P(1 \mid 0)$ and $q = P(0 \mid 1)$. The stationarity condition $\pi^* P = \pi^*$ is equivalent to the flow balance $\pi_1 q = \pi_0 p$, that is,

$$q = rp, \qquad r := \pi_0/\pi_1.$$

The constraint set is therefore described by the single parameter p, and for n states it is an $(n-1)^2$-dimensional polytope (Appendix B.1). Labeling the state with the smaller stationary probability as 0 gives $r \le 1$, so p ranges over all of $[0, 1]$. The smaller r is, the less uniform the stationary distribution is.

Prior and cross-row posterior coupling. We take $p \sim \mathrm{Beta}(\beta, \beta)$ and set $q = rp$ (see Appendix B.1 for n states). If the two rows independently follow $\mathrm{Beta}(\beta, \beta)$, the Bayes predictor is add-β, which uses only (n_{00}, n_{01}) to estimate row 0. The constraint makes the counts (n_{10}, n_{11}) of row 1 also enter the posterior of p (Theorem 2), and we call this phenomenon cross-row posterior coupling.

3.2 Strict suboptimality and closed form

Denote the constrained Bayes predictor by f^*, and let $\kappa_1 = \beta + n_{01} + n_{10}$ and $\kappa_0 = \beta + n_{00}$.

Theorem 1 (Strict suboptimality of add-β for any finite number of states)**.** *Let $n \ge 2$ and suppose that (A1) and (A2) hold and that the stationarity constraint is strictly active, that is, the constraint set $\mathcal{C}(\pi^*)$ is strictly smaller than the unconstrained product of simplices. Then $\mathcal{L}(f_\beta) - \mathcal{L}(f^*) > 0$.*

Remark. The estimator learned by Mamba, as characterized by Bondaschi et al., is therefore no longer optimal on constrained data. Figures 2(a) to (c) show the per-step gap $\mathcal{L}(f_\beta) - \mathcal{L}(f^*)$ for two states (experiment E1, Appendix B.4) and for general numbers of states and orders. The proof is given in Appendices A.1 and A.2.

Theorem 2 (Closed form of f^* for two states)**.** *Let $n = 2$ and suppose that (A1) and (A2) hold. Then the following two statements hold.*

(i) Given the observed history x_1^t, the posterior density of the free parameter p is proportional to $p^{\kappa_1 - 1}(1-p)^{\kappa_0 - 1}(1-rp)^{n_{11}}$.

(ii) f^* has the closed form

$$f^*(1 \mid 0) = \frac{\kappa_1}{\kappa_1 + \kappa_0} \cdot \frac{{}_2F_1(-n_{11}, \kappa_1 + 1; \kappa_1 + \kappa_0 + 1; r)}{{}_2F_1(-n_{11}, \kappa_1; \kappa_1 + \kappa_0; r)}, \qquad f^*(1 \mid 1) = 1 - r\,\mathbb{E}[p \mid x_1^t].$$

Remark. Since κ_1 contains $n_{01} + n_{10}$, f^* also uses the counts from state 1 when the current state is 0, which is the explicit form of cross-row posterior coupling. The hypergeometric functions in the closed form are polynomials of degree n_{11} in r and can be computed directly. The proof is given in Appendix A.3.

3.3 Asymptotic order of the gap

Single strategy. Both predictors converge to the true value, and they differ in the amount of information about the parameter obtained at each step. The gap therefore vanishes at rate $O(1/t)$ (Appendix B.3), and Proposition 1 gives the coefficient of its leading term.

Proposition 1 (Asymptotic order of the gap for any finite number of states)**.** *Let $n \geq 2$, and let the true kernel P° lie in the relative interior of $\mathcal{C}(\pi^*)$ with all entries strictly positive. Suppose that (A1) and (A2) hold and that the prior density is continuous and positive at P°. Then the following two statements hold.*

(i) The cumulative excess risks of add-β and f^* are $\frac{n(n-1)}{2}\log T + O(1)$ and $\frac{(n-1)^2}{2}\log T + O(1)$, respectively, and the cumulative gap is

$$\sum_{t \leq T} \big(\mathrm{excess}_t(f_\beta) - \mathrm{excess}_t(f^*)\big) = \frac{n-1}{2}\log T + O(1).$$

(ii) If the excess risk at step t has regular asymptotics, that is, $t \cdot \mathrm{excess}_t$ converges, then the leading order of the gap at step t is $(n-1)/(2t)$.

Remark. The cumulative excess risk is asymptotically equal to half the parameter dimension times $\log T$. The stationarity constraint reduces the parameter dimension from $n(n-1)$ to $(n-1)^2$, so the leading term of the gap is $\frac{n-1}{2}\log T$. The proof is given in Appendix A.4, and numerical confirmation is shown in Figure 2(d).

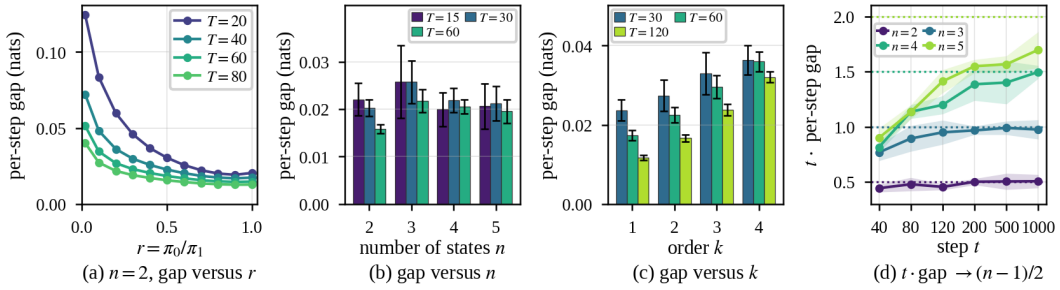


Figure 2: Estimation gap. (a) Per-step gap for two states as a function of r and T. (b) As a function of the number of states n. (c) As a function of the order k. (d) The product of the gap at step t and t tends to $(n-1)/2$.

Strategy switching. After each switch the counts accumulated so far no longer apply, so the gap does not vanish over time. A dwell interval is the period between two consecutive switches. At its beginning the counts are zero, the prediction of f^* equals the corresponding row of the prior mean kernel, and the prediction of add-β is a uniform row. The gap at this point is called the structural component.

Theorem 3 (Characterization of the structural component for any finite number of states)**.** *Let $n \geq 2$ and suppose that (A1) holds. Let Π be any prior supported on $\mathcal{C}(\pi^*)$, let $\bar{P} = \mathbb{E}_\Pi[P]$, and let U be the kernel whose rows are all uniform. Then at zero counts, for any $\beta > 0$, the difference between the expected excess risks of add-β and f^* is*

$$g_{\mathit{struct}} = \sum_i \pi_i^* \, D_{\mathrm{KL}}\big(\bar{P}_{i\cdot} \,\big\|\, U_{i\cdot}\big) \;\geq\; \log n - H(\pi^*),$$

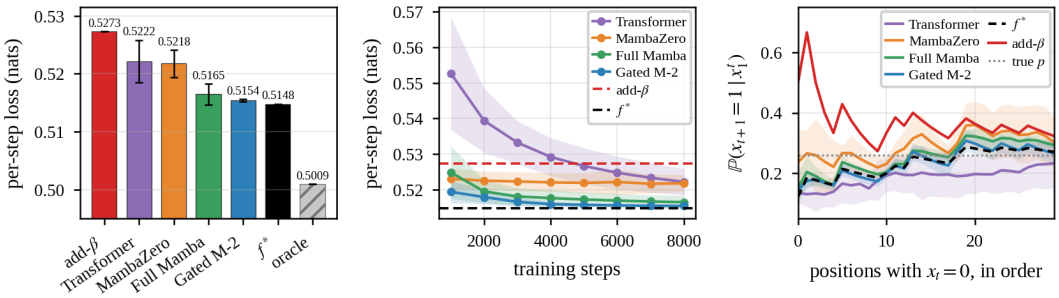
where equality holds if and only if every row of $\bar{P}$ equals π^*. In particular, $g_{\mathit{struct}} > 0$ when π^* is nonuniform.

Remark. The structural component comes only from the known π^*, which add-β has no means to exploit at zero counts. The fraction of steps with zero counts is p_{switch}, and the per-step gap tends to g_{struct} as $p_{\text{switch}} \to 1$ (Appendix B.3, Figure B.1). The proof is given in Appendix A.5.

3.4 Closing the estimation gap

f^* is likewise a function of the transition counts and π^* (Theorem 2), so a sufficiently wide Mamba can represent it in principle, and whether the gap is closed depends on which solution training converges to. Both limited capacity and the inductive bias of convolutional counting could keep training near add-β. Experiment E2 tests these two expectations. It trains the four architectures on constrained single-strategy data ($n = 2$, $r = 0.5$, $\beta = 1$, $T = 120$) for 8000 steps and evaluates them on the same set of 5000 held-out sequences.

After training, the average loss of all four architectures is below that of add-β, so neither expectation holds, but the architectures differ greatly in how far they move toward f^* (Figure 3(a)(b)). Gated M-2 closes about 95% of the gap, whereas the MambaZero variant and the Transformer close only about 40%. On a held-out sequence selected by a rule that depends only on f^*, add-β and the true parameter, the per-step predictions of Gated M-2 and full Mamba stay close to f^* and far from add-β (Figure 3(c)). The stationarity constraint moves the population minimizer of the log loss from add-β to f^*, and the convergence to add-β observed by Bondaschi et al. corresponds to the unconstrained case. The estimation gap is therefore a capability gap, and it decreases as model capacity and the number of training steps increase (Appendix Figure C.9).



(a) loss after 8000 steps (b) loss versus training steps (c) predictions on one sequence

Figure 3: Closing of the estimation gap by the four architectures in E2. (a) Per-step loss after 8000 training steps. (b) Per-step loss as a function of the number of training steps. (c) Next-step predictions at the positions with $x_t = 0$ on one held-out sequence.

4 Representation of the constrained predictor

4.1 Readout analysis and impossibility

Like add-β, f^* depends only on the transition counts (Theorem 2), so the count state in the construction of Bondaschi et al. is also sufficient to represent f^*, and the two differ only at the readout (Figure 1(c)). The hidden states of the three trained Mamba architectures also store the four transition counts linearly. The coefficient of determination of linear decoding is no lower than 0.95 from $t = 20$ to $t = 119$, whereas untrained models with the same initialization reach only 0.46 to 0.67 at $t = 119$ (Probe 2, Appendix Figure C.2). The readout in the construction of Bondaschi et al. reads only the counts of the current row, which suffices exactly for add-β, whereas f^* needs all counts. Whether the constrained optimum can be reached therefore depends on the readout.

Theorem 4 (Upper bound for within-row readouts)**.** *Let $\mathcal{F}_1$ be the class of predictors whose input at state i is only the counts of row i, and let $\mathcal{F}_2$ be the class of predictors whose input is the counts of row i and the position t. Then the Bayes-optimal predictors in the two classes are $f_{\mathcal{F}_1} = \mathbb{E}[f^* \mid x_t,$ the counts of row $x_t]$ and $f_{\mathcal{F}_2} = \mathbb{E}[f^* \mid x_t,$ the counts of row $x_t,\ t]$, respectively. Under assumptions (A1) and (A2), $\mathcal{L}(f_{\mathcal{F}_1}) > \mathcal{L}(f^*)$, that is, the excess risk of any readout whose input is only the within-row counts is strictly higher than that of f^*.*

Remark. The row-selective readout of Bondaschi et al. belongs to $\mathcal{F}_1$, and reaching f^* requires a cross-row readout. For $n = 2$, flow balance gives $|n_{01} - n_{10}| \leq 1$, so the position can almost substitute for cross-row information, and the closed fraction of $f_{\mathcal{F}_2}$ is close to that of f^* (Figure 4(a)). The proof is given in Appendix A.6.

Cross-row information for three states (Probes 5 and 4). For $n = 3$, flow balance gives only two constraints, and the position can no longer substitute for cross-row information. On a family of sequences in which the length and the counts of the current row are both fixed, $f_{\mathcal{F}_2}$ stays unchanged, whereas the prediction of trained full Mamba varies with the counts of the other rows in the same direction as f^* (Figure 4(b), Appendix C.2). Modifying the hidden state along the directions that encode the counts of the other rows reproduces this change, with a magnitude of about 0.86 times the change in f^*, whereas untrained models show no such response (Appendix Figure C.4).

Conjecture 1 (For $n \geq 3$ position cannot substitute for cross-row information)**.** *When $n \geq 3$, $\mathcal{L}(f_{\mathcal{F}_2}) > \mathcal{L}(f^*)$ under assumptions (A1) and (A2).*

Remark. The sequence family of Probe 5 provides numerical evidence, since $f_{\mathcal{F}_2}$ stays unchanged on it whereas f^* varies with the counts of the other rows (Figure 4(b)).

4.2 Constructive representation and width lower bound

A log-linear softmax readout whose input is all counts can represent f^* exactly. The count vector $\mathbf{n} = (n_{00}, n_{01}, n_{10}, n_{11})$ is a sufficient statistic for the constrained posterior, and the posterior density is proportional to $\omega(p)\,\phi_{\mathbf{n}}(p)$, where $\omega(p) = p^{\beta-1}(1-p)^{\beta-1}$ and $\phi_{\mathbf{n}}(p) = p^{n_{01}+n_{10}}(1-p)^{n_{00}}(1-rp)^{n_{11}}$ is a polynomial of degree $t-1$. Since $\log \phi_{\mathbf{n}}(p)$ is an affine function of $\mathbf{n}$, Gauss–Jacobi quadrature turns the posterior mean exactly into a softmax-weighted average of node values.

Theorem 5 (Constructive exact representation of f^*)**.** *Let $n = 2$, $\beta > 0$, and let the horizon be T. Take the L-point Gauss–Jacobi quadrature nodes $\{p_g\}_{g=1}^{L}$ and weights $\{\omega_g\}_{g=1}^{L}$ of the weight function ω on $[0,1]$, where $L = \lceil (T+1)/2 \rceil$. Consider the model whose state layer and readout layer are as follows. The state layer is the SSM recurrence $\mathbf{N}_t = \mathbf{N}_{t-1} + e_{x_{t-1}} e_{x_t}^{\top}$ that accumulates the transition count matrix exactly in the hidden state ($\mathbf{N}_t$ is the transition count matrix), its decay coefficient is constantly 1, the previous symbol is supplied by a convolution with window 2, and e_x is the one-hot vector of symbol x. This is the state construction of Theorem 1 of Bondaschi et al.. The readout layer is*

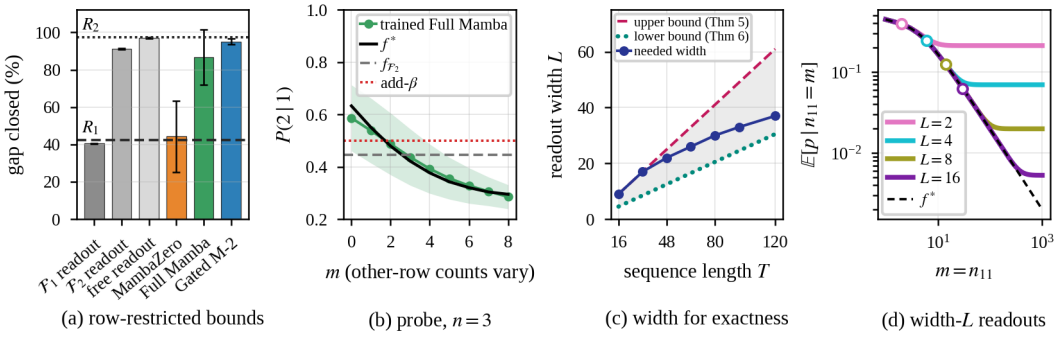
$$\ell_g = \log \omega_g + (n_{01}+n_{10}) \log p_g + n_{00} \log(1-p_g) + n_{11} \log(1-rp_g), \qquad \widehat{p} = \sum_{g=1}^{L} \mathrm{softmax}(\ell)_g \, p_g,$$

where $\widehat{p}$ is the quadrature value of $\mathbb{E}[p \mid x_1^t]$, and the output is $\mathbb{P}_\theta(x_{t+1} = 1 \mid x_1^t) = (1 - x_t)\,\widehat{p} + x_t\,(1 - r\widehat{p})$. Then for every sequence of length at most T, the output of this model is identical to f^*.

Remark. Theorem 5 shares the same state layer with the representation theorem of Bondaschi et al.. The readout of add-β is the normalization of the counts of the current row, whereas the readout of f^* is a softmax-weighted average over log-linear logits. A readout of width L is exact on the all-ones sequence up to exactly $m = 2L - 2$ (Figure 4(d)). The extension to general numbers of states and the proof are given in Appendix A.7.

Theorem 6 (Width lower bound for softmax affine readouts)**.** *Let $n = 2$ and $\beta = 1$. Suppose that the readout belongs to the class described in Theorem 5, that is, its logits are affine functions of the count vector, its output is a linear combination of softmax weights and fixed node values, and its width is L. If this readout is identical to f^* on every sequence of length at most T, then $L \geq (T+2)/4$. In particular, no readout of this class with finite width is identical to f^* on every horizon.*

Remark. The readout width required to represent f^* exactly is therefore $\Theta(T)$ (Figure 4(c)). The lower bound arises because $\mathbb{E}[p \mid n_{11} = m]$ on the all-ones sequence decays at rate $1/m$ (Figure 4(d)), whereas a ratio of finitely many exponential sums can decay only at a constant or exponential rate. The proof is given in Appendix A.7.



(a) row-restricted bounds (b) probe, $n = 3$ (c) width for exactness (d) width-L readouts

Figure 4: Readout structure. (a) The upper bounds R_1 and R_2 for within-row readouts (the closed fractions of $f_{\mathcal{F}_1}$ and $f_{\mathcal{F}_2}$), and the closed fractions of the learned readout and the three Mamba architectures. (b) Three-state probe. (c) The readout width required for exact representation and the two bounds. (d) Readouts of width L on the all-ones sequence.

5 Statistical Limit Under Hidden Switching

5.1 The Residual Loss Is Strictly Positive

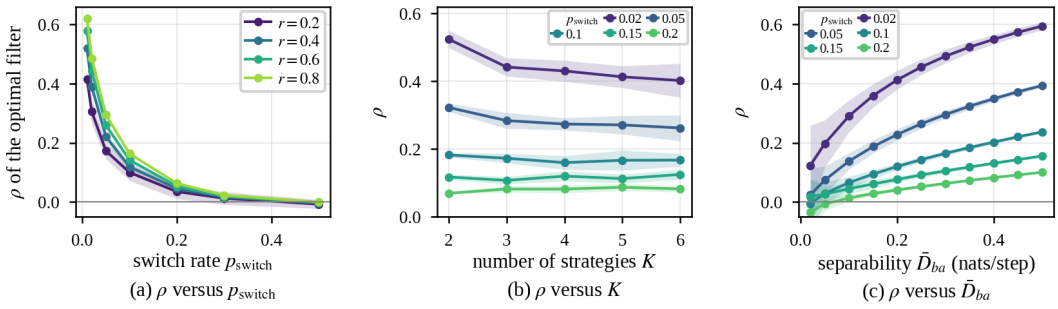
When $\mathcal{P}$ and A are known, the Bayes-optimal predictor in the switching setting is given by the HMM forward filter, which we call the optimal filter.

Lemma 1 (Instantaneous independence)**.** *Suppose that (A1) and (A3) hold and that the initial condition is as specified in Section 2.1. Then for every $t \geq 1$, X_t and Z_t are independent, with $X_t \sim \pi^*$ and $Z_t \sim w$.*

Remark. The current state therefore carries no information about the current strategy, and hence the loss of the oracle predictor is exactly H_{within}. The corollary and its proof are given in Appendix A.8.

Theorem 7 (Statistical limit faced by the optimal filter)**.** *Suppose that (A1) and (A3) hold, and that the strategy set $\mathcal{P}$ contains at least two transition kernels that differ in at least one row i with $\pi_i^* > 0$. Then the loss of the optimal filter satisfies $\mathcal{L}^* > H_{within}$, or equivalently $\rho < 1$ holds strictly.*

Remark. Even with infinite memory and knowledge of all strategy kernels, the loss of the optimal filter is strictly higher than that of the oracle predictor, and every trained model is subject to this limit. The inequality $\rho < 1$ comes from the uncertainty in predicting switches and from the detection delay (Proposition 2, Appendix B.5). The value of ρ for the optimal filter decreases as p_{switch} increases and increases as the strategy separability $\bar{D}_{ba}$ increases (Figure 5). The proof is given in Appendix A.9.



(a) ρ versus p_{switch} (b) ρ versus K (c) ρ versus $\bar{D}_{ba}$

Figure 5: Degree of identification ρ of the optimal filter. (a) As a function of the switching rate. (b) As a function of the number of strategies. (c) As a function of the separability of strategies.

5.2 Lower Bound on Detection Delay

Proposition 2 (Wald-type lower bound on detection delay)**.** *Suppose that (A3) holds, that every row of every strategy kernel is strictly positive, and that the false-alarm probability and the missed-detection probability are both at most $\alpha < 1/2$. Then after the strategy switches from a to b,*

the number of steps τ_{det} from the occurrence of the switch to the completion of detection in any identification procedure satisfies

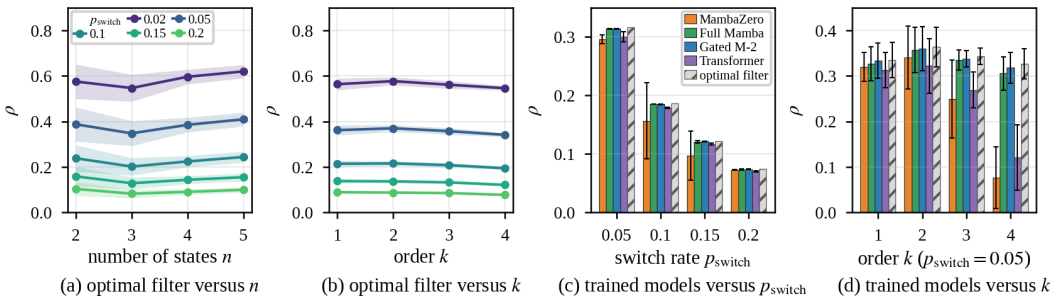
$$\mathbb{E}_b[\tau_{\mathrm{det}}] \geq \frac{\log(1/\alpha)}{\bar{D}_{ba}}\big(1 - o(1)\big), \qquad \bar{D}_{ba} = \mathbb{E}_{X\sim\pi^*}\Big[D_{\mathrm{KL}}\big(P^{(b)}(\cdot \mid X) \,\big\|\, P^{(a)}(\cdot \mid X)\big)\Big].$$

Remark. The closer the two strategy kernels are, the longer the detection delay, which gives the time scale of the loss difference in Theorem 7. The proof is given in Appendix A.10.

Extension to higher orders. If every strategy kernel is of order k, all share the context stationary distribution μ^* (Appendix B.2), and the context chains are irreducible, then with $C_t = (X_{t-k+1}, \ldots, X_t)$ as the context, Lemma 1, Theorem 7 and Proposition 2 hold in parallel (the proof is given in Appendix A.11, and numerical results are shown in Figure 6(b)(d)).

5.3 Constraint imposed by the statistical limit

Experiment E3 trains the four architectures on hidden switching data ($K = 2$, $p \in \{0.15, 0.6\}$, $r = 0.6$, $T = 120$, switching rates from 0.05 to 0.2) for 8000 steps and compares them with the optimal filter on 5000 sequences. For every seed, ρ is below that of the optimal filter, and full Mamba and Gated M-2 almost coincide with it (Figure 6(c)). On switching data of order $k = 1$ to 4, full Mamba and Gated M-2 stay close to the optimal filter at every order, the MambaZero variant and the Transformer fall behind as the order increases, and none of the four architectures exceeds the optimal filter (Figure 6(d), Appendix C.4).



(a) optimal filter versus n (b) optimal filter versus k (c) trained models versus p_{switch} (d) trained models versus k

Figure 6: Extensions of the degree of identification ρ and the trained models. (a) Optimal filter as a function of the number of states. (b) Optimal filter as a function of the order. (c) The four architectures trained for 8000 steps in E3 and the optimal filter. (d) The four trained architectures and the optimal filter at each order.

6 Conclusion

This paper distinguished two problems, estimating the constrained transition kernel and identifying the current transition kernel. Laplacian smoothing is strictly suboptimal under the stationarity constraint, and its difference from the constrained Bayes predictor lies in the readout rather than in the state. Trained Mamba uses the counts of all rows at the readout and closes most of the estimation gap. The residual loss of the optimal filter, in contrast, is strictly positive for any number of states and any order, and trained models can only approach this statistical limit. Therefore, against well-trained sequence models, provable unpredictability must rest on the statistical limit created by hidden switching rather than on the difficulty of estimation.

Limitations. The exact representation result and the strict suboptimality for order k are proved only for two states, and the statement that the position cannot substitute for cross-row information for three or more states remains a conjecture. The training experiments are limited to synthetic data with two to five states and to small models with width at most 32.

AI USE STATEMENT

In this work, we used generative AI tools to polish the language of the manuscript. We have reviewed all AI-assisted text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

E. Akyürek, B. Wang, Y. Kim, and J. Andreas. In-context language learning: Architectures and algorithms. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, 2024.

C. D. Alvarenga, N. Basilico, and S. Carpin. Time-varying graph patrolling against attackers with locally limited and imperfect observation models. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4869–4876, 2019.

A. R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

M. Bondaschi, N. Rajaraman, X. Wei, R. Pascanu, C. Gulcehre, M. Gastpar, and A. V. Makkuva. From Markov to Laplace: How Mamba in-context learns Markov chains. In *International Conference on Learning Representations (ICLR)*, 2026.

X. Cheng, J. Shi, and M. Sha. Cracking channel hopping sequences and graph routes in industrial TSCH networks. *ACM Transactions on Internet Technology*, 20(3):23:1–23:28, 2020.

B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

I. Csiszár and P. C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, 2004.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

F. D'Angelo and N. Flammarion. Transformers learn latent mixture models in-context via mirror descent. In *International Conference on Learning Representations (ICLR)*, 2026.

T. Dao and A. Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 10041–10071, 2024.

E. Edelman, N. Tsilivis, B. L. Edelman, E. Malach, and S. Goel. The evolution of statistical induction heads: In-context learning Markov chains. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. 64273–64311, 2024.

C. Ekbote, A. V. Makkuva, M. Bondaschi, N. Rajaraman, M. Gastpar, J. D. Lee, and P. P. Liang. What one cannot, two can: Two-layer transformers provably represent induction heads on any-order Markov chains. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025.

Federal Communications Commission. 47 CFR §15.247, operation within the bands 902–928 MHz, 2400–2483.5 MHz, and 5725–5850 MHz. Code of Federal Regulations, Title 47, Chapter I, revised as of October 1, 2025, 2025.

R. Grazzi, J. Siems, S. Schrodi, T. Brox, and F. Hutter. Is Mamba capable of in-context learning? In *Proceedings of the Third International Conference on Automated Machine Learning (AutoML)*, volume 256 of *PMLR*, 2024.

A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*, 2023.

J. T. Halloran, M. Gulati, and P. F. Roysdon. Mamba state-space models are Lyapunov-stable learners. *Transactions on Machine Learning Research (TMLR)*, 2025.

Y. Hao, A. Orlitsky, and V. Pichapati. On learning Markov chains. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 648–657, 2018.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991.

J. Hu, Q. Liu, and C. Jin. On limitation of transformer for learning HMMs. *arXiv:2406.04089*, 2024.

M. E. Ildiz, Y. Huang, Y. Li, A. S. Rawat, and S. Oymak. From self-attention to Markov models: Unveiling the dynamics of generative transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, 2024.

T. L. Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929, 1998.

S. Lepage, J. Mary, and D. Picard. Markov chain estimation with in-context learning. *arXiv:2508.03934*, 2025.

G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

A. V. Makkuva, M. Bondaschi, C. Ekbote, A. Girish, A. Nagle, H. Kim, and M. Gastpar. Local to global: Learning dynamics and effect of initialization for transformers. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

A. V. Makkuva, M. Bondaschi, A. Girish, A. Nagle, M. Jaggi, H. Kim, and M. Gastpar. Attention with Markov: A curious case of single-layer transformers. In *International Conference on Learning Representations (ICLR)*, 2025.

N. Muca Cirone, A. Orvieto, B. Walker, C. Salvi, and T. Lyons. Theoretical foundations of deep selective state-space models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

E. Nichani, A. Damian, and J. D. Lee. How transformers learn causal structure with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 38018–38070, 2024.

J. Oh, W. Huang, and T. Suzuki. Mamba can learn low-dimensional targets in-context via test-time feature learning. *arXiv:2510.12026*, 2025.

C. Olsson, N. Elhage, N. Nanda, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. arXiv:2209.11895.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 1954.

J. Park, J. Park, Z. Xiong, N. Lee, J. Cho, S. Oymak, K. Lee, and D. Papailiopoulos. Can Mamba learn how to learn? A comparative study on in-context learning tasks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 39793–39812, 2024.

G. Pólya and G. Szegő. *Problems and Theorems in Analysis II*. Springer, 1976.

N. Rajaraman, M. Bondaschi, K. Ramchandran, M. Gastpar, and A. V. Makkuva. Transformers on Markov data: Constant depth suffices. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024a.

N. Rajaraman, J. Jiao, and K. Ramchandran. An analysis of tokenization: Transformers under Markov data. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024b.

G. Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations (ICLR)*, 2024.

C. Sanford, D. Hsu, and M. Telgarsky. One-layer transformers fail to solve the induction heads task. *arXiv:2408.14332*, 2024.

A. S. Shai, S. E. Marzen, L. Teixeira, A. Gietelink Oldenziel, and P. M. Riechers. Transformers represent belief state geometry in their residual stream. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

V. Tchakaloff. Formules de cubatures mécaniques à coefficients non négatifs. *Bulletin des Sciences Mathématiques (2)*, 81:123–134, 1957.

Y. Tian, N. M. Sushma, H. Mestha, N. Colombo, D. Kappel, and A. Subramoney. Learning in the recurrent state: Gradient descent with linear recurrent networks. *arXiv:2410.11687v3*, 2026.

M. Tiloca, D. De Guglielmo, G. Dini, G. Anastasi, and S. K. Das. DISH: DIstributed SHuffling against selective jamming attack in IEEE 802.15.4e TSCH networks. *ACM Transactions on Sensor Networks*, 15(1):3:1–3:28, 2019.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117–186, 1945.

F. Wang, J. Sun, J. Hu, X. Luo, and T. Wang. A dual-layer evaluation framework for USV maritime patrol based on time-varying Markov chains: Unpredictability index design and adversarial verification. *Ocean Engineering*, 362:126408, 2026.

S. Wang, T. T.-L. Tse, and J. Zhu. Understanding in-context learning beyond transformers: An investigation of state space and hybrid architectures. *arXiv:2510.23006*, 2025.

Q. Xie and A. R. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory*, 43(2):646–657, 1997.

R. Zhou, C. Tian, and S. Diggavi. An information-theoretic approach to understanding transformers' in-context learning of variable-order Markov chains. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026. arXiv:2410.05493.

INDEX OF FIGURES AND TABLES

Figures in the main text.
Figure 1, Framework of this paper (Section 1).
Figure 2, Estimation gap (Section 3.3).
Figure 3, Closing of the estimation gap by the four architectures in E2 (Section 3.4).
Figure 4, Readout structure (Section 4.2).
Figure 5, Degree of identification ρ of the optimal filter (Section 5.1).
Figure 6, Extensions of the degree of identification ρ and the trained models (Section 5.3).

Figures in the appendix.
Figure B.1, Estimation gap under hidden switching (Appendix B.3).
Figure C.1, Prediction trajectories and per-position loss of the models trained in E2 on held-out sequences (Appendix C.1).
Figure C.2, Linear decoding of the transition counts from the hidden state (Appendix C.2).
Figure C.3, Zero-count predictions of the trained model for three states (Appendix C.2).
Figure C.4, State patching of cross-row counts (Appendix C.2).
Figure C.5, Ablation of the readout structure and the capability gap (Appendix C.3).
Figure C.6, Degree of identification of the four architectures in E3 under three training budgets (Appendix C.4).
Figure C.7, Degree of identification of the models trained in E3 as a function of the architecture, the switching rate and the number of training steps (Appendix C.4).
Figure C.8, Closed fraction after training across orders and convolution windows (Appendix C.5).
Figure C.9, Closed fraction after training across numbers of states, widths and numbers of training steps (Appendix C.5).

Tables in the appendix.
Table F.1, Symbols used throughout the paper (Appendix F).

A PROOFS OF THEOREMS AND PROPOSITIONS

A.0 PRELIMINARY LEMMAS

Lemma A.0.1 (Loss decomposition)**.** *Let $f_t^*(\cdot) = \mathbb{P}(X_{t+1} = \cdot \mid X_1^t)$ be the Bayes posterior predictive, and let f_t be any predictor measurable with respect to the history X_1^t (taking values in the probability simplex). Then*

$$\mathbb{E}[-\log f_t(X_{t+1})] = \mathbb{E}[H(f_t^*)] + \mathbb{E}[D_{\mathrm{KL}}(f_t^* \,\|\, f_t)].$$

Hence $\mathcal{L}(f) - \mathcal{L}(f^*) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[D_{\mathrm{KL}}(f_t^* \,\|\, f_t)] \geq 0$, with equality if and only if $f_t = f_t^*$ almost surely for every $t \leq T$.

Proof. By the tower property of conditional expectation, $\mathbb{E}[-\log f_t(X_{t+1})] = \mathbb{E}\big[\mathbb{E}[-\log f_t(X_{t+1}) \mid X_1^t]\big] = \mathbb{E}\big[\sum_j f_t^*(j)(-\log f_t(j))\big]$, and the cross entropy decomposes as $H(f_t^*) + D_{\mathrm{KL}}(f_t^* \,\|\, f_t)$. Nonnegativity and the equality condition follow directly from the properties of the KL divergence. □

Lemma A.0.2 (One-dimensional likelihood-ratio order, strict version)**.** *Let η_1 be a probability measure on $[0,1]$, let $\varphi \geq 0$ be nonincreasing and not η_1-almost everywhere constant, and let $0 < \mathbb{E}_{\eta_1}[\varphi] < \infty$. Let $d\eta_2 \propto \varphi \, d\eta_1$. Then $\mathbb{E}_{\eta_2}[p] < \mathbb{E}_{\eta_1}[p]$.*

Proof. We have $\mathbb{E}_{\eta_2}[p] - \mathbb{E}_{\eta_1}[p] = \mathrm{Cov}_{\eta_1}(p, \varphi)/\mathbb{E}_{\eta_1}[\varphi]$. Moreover,

$$\mathrm{Cov}_{\eta_1}(p, \varphi) = \frac{1}{2} \iint (p - p')\big(\varphi(p) - \varphi(p')\big)\, \eta_1(dp)\, \eta_1(dp'),$$

and the integrand is everywhere ≤ 0 (opposite monotonicity). Let φ_+ and φ_- denote the essential supremum and essential infimum of φ under η_1, so that $\varphi_+ > \varphi_-$, and take $c = (\varphi_+ + \varphi_-)/2$. The sets $\{\varphi > c\}$ and $\{\varphi < c\}$ both have positive η_1-measure, and by monotonicity every point of the former is strictly smaller than every point of the latter. The integrand is therefore strictly < 0 on a product set of positive $\eta_1 \otimes \eta_1$-measure, and the covariance is strictly negative. □

A.1 Proof of Theorem 1 ($n = 2$)

Let $r \in (0, 1]$ and $\beta > 0$ (the prior is nondegenerate). By Lemma A.0, it suffices to show that there exists t with $\mathbb{P}(f_{\beta,t} \neq f^*_t) > 0$.

Step 1 (strict monotonicity). Let ψ_m and $\mathbb{E}[p \mid \kappa_1, \kappa_0, m]$ denote the posterior density and the posterior mean of p when $n_{11} = m$. Since $\psi_{m+1} \propto (1 - rp)\psi_m$, the function $\varphi(p) = 1 - rp$ is strictly decreasing on $(0, 1)$ ($r > 0$), and the posterior is nondegenerate (its density is strictly positive on $(0, 1)$), Lemma A.0 gives $\mathbb{E}[p \mid \kappa_1, \kappa_0, m + 1] < \mathbb{E}[p \mid \kappa_1, \kappa_0, m]$. That is, the posterior mean is strictly decreasing in n_{11}.

Step 2 (a pair of histories with positive probability). Fix $c \geq 1$, and for $m \geq 0$ define the path $\mathcal{H}_m$ as follows. It starts with $X_1 = 0$, followed by c transitions $0 \to 0$, one transition $0 \to 1$, m transitions $1 \to 1$ and one transition $1 \to 0$. The current state is 0, and the counts are $(n_{00}, n_{01}, n_{10}, n_{11}) = (c, 1, 1, m)$. The probability of this path under the prior is $\mathbb{E}\big[\pi^*_0\, p(1 - p)^c\, (rp)(1 - rp)^m\big] > 0$ (the integrand is almost surely strictly positive for $p \in (0, 1)$).

Step 3 (conclusion). Take $m \neq m'$. The predictor f_β takes the same value on $\mathcal{H}_m$ and $\mathcal{H}_{m'}$ (it depends only on $(n_{00}, n_{01}) = (c, 1)$ and the current state), whereas f^* takes strictly different values by Step 1. Hence $f^*_t \neq f_{\beta,t}$ on at least one of the two histories of positive probability. Two different binary distributions have strictly positive KL divergence, and substituting into Lemma A.0 gives $\mathcal{L}(f_\beta) - \mathcal{L}(f^*) > 0$. □

A.2 Proof of Theorem 1 ($n \geq 3$)

Let $n \geq 3$ (the case $n = 2$ is covered by Appendix A.1) and let π^* lie in the interior. The constraint set $\mathcal{C}(\pi^*)$ is a compact convex polytope, abbreviated as $\mathcal{C}$ below. The kernel P_π whose rows all equal π^* satisfies $\pi^* P_\pi = \pi^*$ and has strictly positive entries, so it lies in the relative interior of $\mathcal{C}$, and the relative interior of $\mathcal{C}$ is therefore nonempty. The prior $\mathcal{Q}$ is the product $\mathrm{Dir}(\beta)$ on the free rows restricted to the feasible region (equivalently, an absolutely continuous measure on $\mathcal{C}$ with a strictly positive density on its relative interior). By absolute continuity, every entry of P is almost surely strictly positive (the entries of the dependent row are affine functions of the free coordinates and vanish only on hyperplanes of measure zero).

Choice of states and the count event. Take $j \in \arg\min_m \pi^*_m$, any $u \neq j$ (so that $\pi^*_j \leq \pi^*_u$), and any $i \notin \{j, u\}$ (possible since $n \geq 3$). Define the event $\mathcal{E}_R$ as follows. The trajectory starts with $X_1 = u$, followed by R round trips $u \to j \to u$ and then one step $u \to i$. The counts are $n_{uj} = n_{ju} = R$ and $n_{ui} = 1$, all other counts are 0, and the current state is i. The probability of this event under the prior is $\mathbb{P}(\mathcal{E}_R) = \mathbb{E}\big[\pi^*_u\, P^R_{uj} P^R_{ju} P_{ui}\big] > 0$ (the integrand is almost surely strictly positive). On this event $f_\beta(j' \mid i) = \beta/(n\beta) = 1/n$ (all counts of row i are zero), which does not depend on R.

Lemma A.2.1 (Maximum and properties of the maximizer)**.** *Let $\Phi(P) = \log P_{uj} + \log P_{ju}$ (taken as $-\infty$ when a factor is zero). Then $\sup_{\mathcal{C}} \Phi = \log(\pi^*_j/\pi^*_u) =: \Phi^*$, and the supremum is attained. Moreover, for every $P \in \mathcal{C}$,*

$$P_{uj} P_{ju} \;\leq\; \frac{\pi^*_j - \pi^*_i P_{ij}}{\pi^*_u}. \tag{$*$}$$

Proof. The stationarity equation of column j, $\sum_m \pi^*_m P_{mj} = \pi^*_j$, gives $\pi^*_u P_{uj} = \pi^*_j - \sum_{m \neq u} \pi^*_m P_{mj} \leq \pi^*_j - \pi^*_i P_{ij}$ (dropping the remaining nonnegative terms). Multiplying by $P_{ju} \leq 1$ gives equation $*$. In particular ($P_{ij} \geq 0$), $e^{\Phi(P)} \leq \pi^*_j/\pi^*_u$.

For attainability, define $\widehat{P}$ as follows. Row j puts all its mass on u. Row u takes $\widehat{P}_{uj} = \pi^*_j/\pi^*_u$ and $\widehat{P}_{uu} = 1 - \pi^*_j/\pi^*_u$ (nonnegative because $\pi^*_j \leq \pi^*_u$), with all other entries 0. Every other row m takes the self-loop $\widehat{P}_{mm} = 1$. We verify $\pi^* \widehat{P} = \pi^*$ column by column. Column $m \notin \{j, u\}$ gives π^*_m. Column j gives $\pi^*_u(\pi^*_j/\pi^*_u) = \pi^*_j$. Column u gives $\pi^*_j + \pi^*_u - \pi^*_j = \pi^*_u$. Moreover $\Phi(\widehat{P}) = \Phi^*$. □

Corollary A.2.2 (Quantitative inclusion). *For every $\delta \geq 0$, $\{P \in \mathcal{C} : \Phi(P) \geq \Phi^* - \delta\} \subseteq \{P : P_{ij} \leq (\pi_j^*/\pi_i^*)(1 - e^{-\delta})\}$.*

Proof. By equation $*$, $e^{\Phi^* - \delta} \leq (\pi_j^* - \pi_i^* P_{ij})/\pi_u^*$, that is, $\pi_i^* P_{ij} \leq \pi_j^*(1 - e^{-\delta})$. □

Lemma A.2.3 (Posterior concentration). *The posterior given $\mathcal{E}_R$ is $d\mathcal{Q}_R(P) \propto e^{R\Phi(P)} P_{ui}\, d\mathcal{Q}(P)$. For every $\delta > 0$, $\mathcal{Q}_R(\Phi < \Phi^* - \delta) \to 0$ as $R \to \infty$.*

Proof. The numerator satisfies $\int_{\{\Phi < \Phi^* - \delta\}} e^{R\Phi} P_{ui}\, d\mathcal{Q} \leq e^{R(\Phi^* - \delta)}$. The denominator is at least $e^{R(\Phi^* - \delta/2)}\, \zeta(\delta)$, where $\zeta(\delta) = \int_{\{\Phi > \Phi^* - \delta/2\}} P_{ui}\, d\mathcal{Q}$. We now show $\zeta(\delta) > 0$. Take $\widehat{P}_\varepsilon = (1 - \varepsilon)\widehat{P} + \varepsilon P_\pi \in \mathcal{C}$ (by convexity). The function Φ is continuous at $\widehat{P}_\varepsilon$ and $\Phi(\widehat{P}_\varepsilon) \to \Phi^*$ as $\varepsilon \to 0$, so there exists a small ε with $\Phi(\widehat{P}_\varepsilon) > \Phi^* - \delta/2$. Moreover, $\widehat{P}_\varepsilon$ lies in the relative interior of $\mathcal{C}$ (as a convex combination with a relative interior point), and $(\widehat{P}_\varepsilon)_{ui} = \varepsilon \pi_i^* > 0$. By the continuity of Φ and P_{ui}, there is a relatively open neighborhood of $\widehat{P}_\varepsilon$ on which $\Phi > \Phi^* - \delta/2$ and $P_{ui} > \varepsilon \pi_i^*/2$. The prior density is strictly positive on the relative interior, so $\zeta(\delta) > 0$. Hence $\mathcal{Q}_R(\Phi < \Phi^* - \delta) \leq e^{-R\delta/2}/\zeta(\delta) \to 0$. □

By Corollary A.2 and Lemma A.2,

$$\mathbb{E}[P_{ij} \mid \mathcal{E}_R] \;\leq\; \frac{\pi_j^*}{\pi_i^*}(1 - e^{-\delta}) + \mathcal{Q}_R(\Phi < \Phi^* - \delta) \;\xrightarrow{R \to \infty}\; \frac{\pi_j^*}{\pi_i^*}(1 - e^{-\delta})$$

for every $\delta > 0$, so $f_R^*(j \mid i) = \mathbb{E}[P_{ij} \mid \mathcal{E}_R] \to 0$. In contrast, $f_\beta(j \mid i) = 1/n$ is constant. Taking R large enough gives $f^* \neq f_\beta$ on $\mathcal{E}_R$ (which has positive probability). Lemma A.0 then gives a strict gap. □

Remark (Covariance identity). Take row u as the dependent row. Then $P_{uj} = \big(\pi_j^* - \sum_{m \neq u} \pi_m^* P_{mj}\big)/\pi_u^*$ is an affine function of the entries of the free rows, and the coefficient of P_{ij} is $-\pi_i^*/\pi_u^* < 0$. The posterior density contains the likelihood factor $\prod_j P_{uj}^{n_{uj}}$ of row u, which couples the free rows, so the posterior does not factorize over rows. Treating n_{uj} as a continuous parameter, the posterior is an exponential family with $\log P_{uj}$ as sufficient statistic ($\beta > 0$ ensures that $|\log P_{uj}|$ is integrable on the compact set $\mathcal{C}$), and differentiating the normalizing constant gives $\partial\, \mathbb{E}[P_{ij} \mid \text{counts}]/\partial n_{uj} = \mathrm{Cov}_{\text{post}}(P_{ij}, \log P_{uj})$. This covariance is nonzero when the constraint is nondegenerate, so the prediction of f^* in row i depends on the counts of row u, whereas add-β uses only the counts of row i. This coupling is not limited to the limiting construction above.

A.3 Proof of Theorem 2

Likelihood and posterior. The free parameter is $p = P(1 \mid 0)$, and the constraint gives $q = P(0 \mid 1) = rp$. Given p, the likelihood of a trajectory of length t factorizes over transitions. The transitions from state 0 contribute $p^{n_{01}}(1 - p)^{n_{00}}$, and the transitions from state 1 contribute $q^{n_{10}}(1 - q)^{n_{11}} = r^{n_{10}} p^{n_{10}} (1 - rp)^{n_{11}}$. The constant $r^{n_{10}}$ and the initial distribution π^* (which does not depend on p) are absorbed into the normalization. Multiplying by the $\mathrm{Beta}(\beta, \beta)$ prior gives the posterior density

$$\psi(p \mid x_1^t) \propto p^{\kappa_1 - 1}(1 - p)^{\kappa_0 - 1}(1 - rp)^{n_{11}}, \qquad \kappa_1 = \beta + n_{01} + n_{10},\ \kappa_0 = \beta + n_{00},$$

which is integrable on $(0, 1)$ (the product of a polynomial and a Beta kernel).

Posterior mean. Expanding $(1 - rp)^{n_{11}}$ binomially as $\sum_{j=0}^{n_{11}} \binom{n_{11}}{j}(-r)^j p^j$ and computing the Beta integrals term by term gives

$$\int_0^1 p^{\kappa_1 - 1 + \iota}(1 - p)^{\kappa_0 - 1}(1 - rp)^{n_{11}}\, dp = \sum_{j=0}^{n_{11}} \binom{n_{11}}{j}(-r)^j\, B(\kappa_1 + \iota + j,\, \kappa_0), \qquad \iota \in \{0, 1\}.$$

Using $B(\kappa_1 + \iota + j, \kappa_0)/B(\kappa_1 + \iota, \kappa_0) = (\kappa_1 + \iota)_j/(\kappa_1 + \iota + \kappa_0)_j$ and $\binom{m}{j}(-1)^j = (-m)_j/j!$, the two sums are identified as the terminating hypergeometric functions $B(\kappa_1 + \iota, \kappa_0)\, {}_2F_1(-n_{11}, \kappa_1 + \iota; \kappa_1 + \iota + \kappa_0; r)$ (finite sums, with no convergence condition). Taking the ratio and using $B(\kappa_1 + 1, \kappa_0)/B(\kappa_1, \kappa_0) = \kappa_1/(\kappa_1 + \kappa_0)$ gives the formula in the main text. For $x_t = 1$, $f^*(x_t = 1) = \mathbb{E}[1 - q \mid x_1^t] = 1 - r\, \mathbb{E}[p \mid x_1^t]$. □

A.4 Proof of Proposition 1

Each of the two predictors is the Bayes predictor of its own model. add-β corresponds to the product Dirichlet prior with independent rows, whose parameter space is the product of n open simplices, of dimension $n(n-1)$. f^* corresponds to the same prior restricted to $\mathcal{C}(\pi^*)$, whose parameter space is the relative interior of the constraint set, an open set in $\mathbb{R}^{(n-1)^2}$ with the free rows as coordinates. The true kernel P° has strictly positive entries and lies in the relative interior of $\mathcal{C}(\pi^*)$, so it lies in the interior of both parameter spaces.

For a smooth parametric family of finite-state ergodic Markov chains, when the prior density is continuous and positive at the true value and the Fisher information matrix is positive definite at the true value, the cumulative excess risk (cumulative KL redundancy) of the Bayes predictor equals half the parameter dimension times $\log T$ plus $O(1)$. The i.i.d. case of this result is given by Clarke & Barron (1990), the corresponding result for Dirichlet mixtures on finite-state Markov chains is given in Section 6 of Csiszár & Shields (2004), and general smooth parametric families are treated by Barron et al. (1998).

The Fisher information of the unconstrained family at P° is the direct sum of blocks of the form $\pi_i^\circ \operatorname{diag}(1/P_{i\cdot}^\circ)$ over the rows, which is positive definite. The Fisher information of the constrained family is the restriction of this positive definite quadratic form to the constraint subspace, which is still positive definite. The unconstrained family has dimension $n(n-1)$ and the constrained family has dimension $(n-1)^2$, and the difference is $n(n-1) - (n-1)^2 = n-1$. Subtracting the two expansions gives the cumulative gap $\frac{n-1}{2}\log T + O(1)$. If $t \cdot \text{excess}_t$ converges, then since the cumulative excess risk is asymptotically half the parameter dimension times $\log T$, the limit is half the parameter dimension. The difference between the limits of the two predictors is therefore $(n-1)/2$, so the leading order of the gap at step t is $(n-1)/(2t)$. □

A.5 Proof of Theorem 3

The two predictors at zero counts. When no transition has been observed yet, the posterior equals the prior, so the prediction of the constrained Bayes predictor at state i is $\mathbb{E}_\Pi[P_{i\cdot}] = \bar{P}_{i\cdot}$. All counts of add-$\beta$ are zero, and its prediction is $(0+\beta)/(0+n\beta) = 1/n$, that is, $U_{i\cdot}$, which does not depend on β. The state at which the first prediction is made satisfies $X_1 \sim \pi^*$. In the switching setting, the switching times are independent of the state process, and by Lemma 1 the state at a switching time also follows π^*. The weighting by π_i^* is therefore valid in both cases.

Compensation identity. For any fixed prediction row $Q_{i\cdot}$ and the Bayes prediction row $\bar{P}_{i\cdot}$,

$$\mathbb{E}_\Pi\big[D_{\mathrm{KL}}(P_{i\cdot} \,\|\, Q_{i\cdot}) - D_{\mathrm{KL}}(P_{i\cdot} \,\|\, \bar{P}_{i\cdot})\big] = \sum_j \mathbb{E}_\Pi[P_{ij}] \log \frac{\bar{P}_{ij}}{Q_{ij}} = D_{\mathrm{KL}}(\bar{P}_{i\cdot} \,\|\, Q_{i\cdot}),$$

because the entropy terms $\sum_j P_{ij} \log P_{ij}$ of the two KL divergences cancel and the remainder is linear in P. Taking $Q_{i\cdot} = U_{i\cdot}$ and summing with weights π_i^* gives $g_{\text{struct}} = \sum_i \pi_i^* D_{\mathrm{KL}}(\bar{P}_{i\cdot} \,\|\, U_{i\cdot})$.

The prior mean satisfies the constraint. The constraint $\pi^* P = \pi^*$ is linear in P and Π is supported on the constraint set, so $\pi^* \bar{P} = \mathbb{E}_\Pi[\pi^* P] = \pi^*$.

Lower bound. Let $\pi^* \otimes \bar{P}$ denote the joint distribution of (X_t, X_{t+1}) when $X_t \sim \pi^*$ and the transition kernel is $\bar{P}$, and define $\pi^* \otimes U$ analogously. The chain rule for conditional KL divergence gives $\sum_i \pi_i^* D_{\mathrm{KL}}(\bar{P}_{i\cdot} \,\|\, U_{i\cdot}) = D_{\mathrm{KL}}(\pi^* \otimes \bar{P} \,\|\, \pi^* \otimes U)$. Applying the data processing inequality to the marginal of X_{t+1} gives

$$D_{\mathrm{KL}}(\pi^* \otimes \bar{P} \,\|\, \pi^* \otimes U) \;\geq\; D_{\mathrm{KL}}(\pi^* \bar{P} \,\|\, \pi^* U) = D_{\mathrm{KL}}(\pi^* \,\|\, \mathrm{Unif}_n) = \log n - H(\pi^*),$$

where we used $\pi^* \bar{P} = \pi^*$ and $\pi^* U = \mathrm{Unif}_n$. Equality holds in the data processing inequality if and only if the likelihood ratio $\bar{P}_{ij}/U_{ij} = n\bar{P}_{ij}$ does not depend on i, that is, all rows of $\bar{P}$ are equal. Equal rows together with $\pi^* \bar{P} = \pi^*$ imply that every row equals π^*. The right-hand side is strictly positive when π^* is nonuniform. □

A.6 Proofs of Theorem 4 and Proposition B.6

Theorem 4. The log loss is a strictly proper scoring rule, so under any information constraint $\mathcal{I}_t$ (a σ-algebra), the predictor that minimizes the expected loss is the conditional probability $\mathbb{P}(X_{t+1} = \cdot \mid \mathcal{I}_t)$. f^* is the conditional probability given the σ-algebra generated by all counts and the current state, and the σ-algebras corresponding to $\mathcal{F}_1$ and $\mathcal{F}_2$ are sub-σ-algebras of it. By the tower property of conditional expectation, $\mathbb{P}(X_{t+1} = \cdot \mid \mathcal{I}_t) = \mathbb{E}[f^* \mid \mathcal{I}_t]$, which gives $f_{\mathcal{F}_1}$ and $f_{\mathcal{F}_2}$. Strictness is argued as follows. If $\mathcal{L}(f_{\mathcal{F}_1}) = \mathcal{L}(f^*)$, then by the loss decomposition of Lemma A.0, $f_{\mathcal{F}_1} = f^*$ almost surely, that is, f^* is a function of the counts of the current row. However, the proofs in Appendices A.1 and A.2 each give a family of histories of positive probability on which f^* varies with the counts of other rows while the counts of the current row stay fixed, which is a contradiction. □

Proposition B.6. On the bounded count domain $\{0, \dots, T\}^4$, f^* is given by the closed form of Theorem 2 or by its log-domain integral representation, and it is a continuous function of the count vector (any function on finitely many points is continuous, and any continuous extension to $[0, T]^4$ can be taken). Two-layer feedforward networks are uniformly dense in the continuous functions on compact sets (Cybenko, 1989; Hornik, 1991), so there exist parameters for which the uniform deviation is less than ε. The premise that the state maintains the counts linearly ensures that the input of the readout can be obtained from the state by a linear transformation, and this linear transformation is absorbed into the first layer of the readout. □

A.7 Proofs of Theorems 5 and 6

Theorem 5. The count vector is a sufficient statistic for the constrained posterior. By the derivation of Theorem 2, when the current state is 0 the posterior density is proportional to $\omega(p)\,\phi_{\mathbf{n}}(p)$, so

$$\mathbb{E}[p \mid x_1^t] = \frac{\int_0^1 p\,\phi_{\mathbf{n}}(p)\,\omega(p)\,dp}{\int_0^1 \phi_{\mathbf{n}}(p)\,\omega(p)\,dp}.$$

The integrands of the numerator and the denominator, without the weight ω, are polynomials of degree t and $t-1$, respectively. The L-point Gauss–Jacobi rule for the weight function ω is exact for polynomials of degree at most $2L-1$. The choice $L = \lceil (T+1)/2 \rceil$ ensures $2L-1 \geq T \geq t$, so the two integrals equal $\sum_g \omega_g\, p_g\, \phi_{\mathbf{n}}(p_g)$ and $\sum_g \omega_g\, \phi_{\mathbf{n}}(p_g)$, respectively. Since $\phi_{\mathbf{n}}(p_g) = \exp\big((n_{01}+n_{10})\log p_g + n_{00}\log(1-p_g) + n_{11}\log(1-rp_g)\big)$, their ratio is exactly $\sum_g \operatorname{softmax}(\ell)_g\, p_g = \widehat{p}$. The recurrence of the state layer gives the exact counts. When the current state is 1, $f^* = 1 - r\,\mathbb{E}[p \mid x_1^t]$, which is the second term in the output formula of Theorem 5. □

Theorem 6. On the all-ones sequence of length $m+1$, the counts are $(0, 0, 0, m)$ and the current state is 1, so $f^* = 1 - r\,\mathbb{E}[p \mid m]$. For $\beta = 1$, direct integration gives

$$\mu_0(m) = \int_0^1 (1-rp)^m\,dp = \frac{1-(1-r)^{m+1}}{r(m+1)},$$

$$\mu_1(m) = \int_0^1 p\,(1-rp)^m\,dp = \frac{1}{r^2}\left[\frac{1-(1-r)^{m+1}}{m+1} - \frac{1-(1-r)^{m+2}}{m+2}\right],$$

where $\mathbb{E}[p \mid m] = \mu_1(m)/\mu_0(m)$, and $\mathbb{E}[p \mid m] \sim 1/(r(m+2))$ as $m \to \infty$. On this family of sequences, the output of the readout described in Theorem 5 is $F(m) = \sum_g c_g \lambda_g^m / \sum_g d_g \lambda_g^m$, where $\lambda_g = e^{a_g} > 0$, $d_g = e^{b_g} > 0$, $c_g = v_g d_g$, a_g is the coefficient of the logits with respect to n_{11}, v_g is the node value, and b_g is the constant term of the logits. Let $\widetilde{c}_g = (d_g - c_g)/r$. Then $F \equiv f^*$ for $m = 0, \dots, T$ is equivalent to $\widetilde{F}(m) := (1 - F(m))/r = \sum_g \widetilde{c}_g \lambda_g^m / \sum_g d_g \lambda_g^m = \mu_1(m)/\mu_0(m)$, which is again a ratio of exponential sums of the same type. Multiplying both sides by $r^2(m+1)(m+2)$ and rearranging, the equality becomes

$$E(m) := r(m+2)\big(1-(1-r)^{m+1}\big)\sum_g \widetilde{c}_g \lambda_g^m - \Big[(m+2)\big(1-(1-r)^{m+1}\big) - (m+1)\big(1-(1-r)^{m+2}\big)\Big]\sum_g d_g \lambda_g^m = 0.$$

E is an exponential polynomial with bases $\{\lambda_g\}$ and $\{(1-r)\lambda_g\}$ and polynomial coefficients of degree at most 1, and the number of its bases is at most $2L$. If some terms share the same base, merging them reduces the number of bases, and the zero bound below still holds. For an exponential polynomial $\sum_j q_j(x)\gamma_j^x$ with positive real bases that does not vanish identically, the number of real zeros is at most $\sum_j(\deg q_j + 1) - 1$ (Pólya & Szegő, 1976, Volume II, Part Five, Problem 75), which here is at most $4L - 1$. We now show that E does not vanish identically. The softmax definition guarantees $d_g = e^{b_g} > 0$, so the term with the largest base in the denominator of $\widetilde{F}(m)$ has a positive coefficient. As $m \to \infty$, $\widetilde{F}(m)$ therefore tends to a finite constant or decays at an exponential rate, and it cannot decay at rate $1/m$, whereas $\mu_1(m)/\mu_0(m)$ decays at rate $1/m$. If $E \equiv 0$, then $\widetilde{F}(m) = \mu_1(m)/\mu_0(m)$ for all m, which is a contradiction. Hence the $T + 1$ zeros of E at $m = 0, \ldots, T$ satisfy $T + 1 \le 4L - 1$, that is, $L \ge (T + 2)/4$. With finite width the readout cannot be identical to f^* on every horizon, because T can be arbitrarily large. □

Extension to general numbers of states. Replace the Gauss–Jacobi rule on the interval by a quadrature rule on the constraint set $\mathcal{C}(\pi^*)$ with positive weights and algebraic degree of exactness T. The logits are still affine functions of the counts, with the logarithms of the kernel entries at each node as coefficients, and the softmax structure and the count recurrence of the state layer are unchanged. Tchakaloff's theorem (Tchakaloff, 1957) guarantees that such a quadrature rule exists, with at most $\binom{(n-1)^2+T}{(n-1)^2}$ nodes. If the node kernel $P^{(g)}$ has an entry $P^{(g)}_{ij} = 0$, then $\log P^{(g)}_{ij}$ in the corresponding logit is taken as $-\infty$, which is a notational convention rather than an architectural parameter. When $n_{ij} \ge 1$, this logit is $-\infty$ and the softmax weight is zero, consistent with the integrand vanishing at this node. When $n_{ij} = 0$, the coefficient of this term is zero. This paper states the case of general n as an extension and does not list it as a theorem.

Numerical check. When the width satisfies Theorem 5, the constructed model agrees with f^* up to floating-point rounding. A readout of width L is exact on the family of all-ones sequences exactly for $m \le 2L - 2$.

A.8 Proof of Lemma 1 and its corollary

Base case. For $t = 1$, the initial condition of Section 2.1 gives $X_1 \sim \pi^*$ and $Z_1 \sim w$, and the two are independent.

Induction. Suppose that (X_{t-1}, Z_{t-1}) are independent with marginals (π^*, w). By the generating mechanism (Z_t transitions from Z_{t-1} according to A and is conditionally independent of X_{t-1}, and $X_t \sim P^{(Z_t)}(\cdot \mid X_{t-1})$),

$$\begin{aligned}\mathbb{P}(X_t = j, Z_t = \ell) &= \sum_{i,m} \mathbb{P}(X_{t-1} = i, Z_{t-1} = m)\, A_{m\ell}\, P^{(\ell)}(j \mid i) \\ &= \Big[\sum_m w_m A_{m\ell}\Big]\Big[\sum_i \pi^*_i P^{(\ell)}(j \mid i)\Big] = w_\ell\, \pi^*_j.\end{aligned}$$

The second equality uses the induction hypothesis (the joint distribution factorizes) and exchanges the sums. The last equality uses the stationarity of w and the shared stationary distribution $\pi^* P^{(\ell)} = \pi^*$. □

Corollary (independence of (X_t, Z_{t+1})). $\mathbb{P}(X_t = j, Z_{t+1} = \ell) = \sum_m \mathbb{P}(X_t = j, Z_t = m) A_{m\ell} = \pi^*_j \sum_m w_m A_{m\ell} = \pi^*_j w_\ell$. Hence the loss of the oracle predictor, $\mathbb{E}\big[H\big(P^{(Z_{t+1})}(\cdot \mid X_t)\big)\big] = \sum_\ell w_\ell H_{\pi^*}(P^{(\ell)}) = H_{\text{within}}$, holds exactly. For the predictor that uses the strategy-averaged kernel $\bar{P}^w$, the conditional distribution of the next state in its loss $\mathbb{E}[-\log \bar{P}^w(X_{t+1} \mid X_t)]$ is $P^{(Z_{t+1})}(\cdot \mid X_t)$, whose mixture over Z_{t+1} is exactly $\bar{P}^w(\cdot \mid X_t)$ (by independence), so this loss is $H_{\pi^*}(\bar{P}^w)$. The version of order k (with C_t as the context and μ^* as the context stationary distribution) follows in the same way (see Appendix A.11). □

The identity $H_{\text{between}} = I(X_{t+1}; Z_{t+1} \mid X_t)$ also rests on this lemma. Lemma 1 gives independence only at a single time. The state trajectory and the strategy sequence as a whole are dependent, since otherwise filtering based on the history would provide no information.

A.9 Proof of Theorem 7

Optimal predictor. Let the belief be $b_t(\ell) = \mathbb{P}(Z_t = \ell \mid X_1^t)$. By the conditional independence structure of the model,

$$\mathbb{P}(X_{t+1} = \cdot \mid X_1^t) = \sum_\ell \mathbb{P}(Z_{t+1} = \ell \mid X_1^t)\, P^{(\ell)}(\cdot \mid X_t) = \sum_\ell \bar{b}_{t+1}(\ell) P^{(\ell)}(\cdot \mid X_t),$$

where $\bar{b}_{t+1} = A^\top b_t$, which is the HMM forward filter.

Comparison with the oracle predictor. Given (X_1^t, Z_{t+1}), $X_{t+1} \sim P^{(Z_{t+1})}(\cdot \mid X_t)$. For the predictor $Q_t(\cdot) = \sum_\ell \bar{b}_{t+1}(\ell) P^{(\ell)}(\cdot \mid X_t)$, the tower property of conditional expectation gives

$$\mathcal{L}_t^* = \mathbb{E}\big[H\big(P^{(Z_{t+1})}(\cdot \mid X_t)\big)\big] + \mathbb{E}\big[D_{\mathrm{KL}}\big(P^{(Z_{t+1})}(\cdot \mid X_t) \,\big\|\, Q_t\big)\big] = H_{\text{within}} + \mathbb{E}\big[D_{\mathrm{KL}}(\cdot \,\|\, \cdot)\big],$$

where the first term follows from the corollary of Lemma 1 (Appendix A.8).

(i) $\bar{b}_{t+1}$ has full support, since $\bar{b}_{t+1}(\ell) = \sum_m b_t(m) A_{m\ell} \geq \min_m A_{m\ell} > 0$ (the off-diagonal entries of A are > 0 and the diagonal entries are $1 - p_{\text{switch}} > 0$). (ii) Take a row i in which the kernels differ (there exist strategies $a \neq b$ with $P^{(a)}(\cdot \mid i) \neq P^{(b)}(\cdot \mid i)$). The event $\{X_t = i\}$ has probability $\pi_i^* > 0$. On this event, the conditional expected KL divergence given the history is $\sum_\ell \bar{b}_{t+1}(\ell) D_{\mathrm{KL}}(P^{(\ell)}(\cdot \mid i) \,\|\, Q_t)$, which is zero if and only if every ℓ in the support satisfies $P^{(\ell)}(\cdot \mid i) = Q_t$. By full support, this requires $P^{(a)}(\cdot \mid i) = P^{(b)}(\cdot \mid i)$, which contradicts the kernels being different. This conditional expectation is therefore a strictly positive random variable on $\{X_t = i\}$, and its integral over an event of positive probability is strictly positive, that is, $\mathcal{L}_t^* > H_{\text{within}}$ for every t. Since $\mathcal{L}^*$ is the average of $\mathcal{L}_t^*$ over T steps, $\mathcal{L}^* > H_{\text{within}}$, that is, $\rho < 1$. □

A.10 Proof of Proposition 2

Assumptions. After the switch, the observations come from the (context) chain of strategy b with initial distribution μ^* (Lemma 1 ensures this also in the online setting). Every row of every strategy kernel is strictly positive (or $P^{(b)}$ is row-wise absolutely continuous with respect to $P^{(a)}$ with a uniformly bounded log-likelihood ratio, with bound denoted $\Lambda_{\max}$). An identification procedure consists of a stopping time τ (adapted to the observation stream) and a decision $\widehat{Z} \in \{a, b\}$, with both error probabilities $\leq \alpha < 1/2$. Without loss of generality, assume $\mathbb{E}_b[\tau] < \infty$ (otherwise the lower bound holds trivially).

Step 1 (lower bound on the information). Let $\Lambda_\tau = \sum_{t<\tau} \lambda_t$ with $\lambda_t = \log \frac{P^{(b)}(X_{t+1}|C_t)}{P^{(a)}(X_{t+1}|C_t)}$. The KL divergence between the two measures on the stopped σ-algebra $\mathcal{G}_\tau$ is exactly $\mathbb{E}_b[\Lambda_\tau]$ (the likelihood-ratio identity at a stopping time). The decision $\widehat{Z}$ is measurable with respect to $\mathcal{G}_\tau$, and the data processing inequality gives

$$\mathbb{E}_b[\Lambda_\tau] = D_{\mathrm{KL}}\big(\mathbb{P}_b|_{\mathcal{G}_\tau} \,\big\|\, \mathbb{P}_a|_{\mathcal{G}_\tau}\big) \;\geq\; D_{\mathrm{KL}}\big(\mathrm{Law}_b(\widehat{Z}) \,\big\|\, \mathrm{Law}_a(\widehat{Z})\big) \;\geq\; \mathrm{kl}(1-\alpha, \alpha),$$

where $\mathrm{kl}(\cdot, \cdot)$ is the binary KL divergence. As $\alpha \to 0$, $\mathrm{kl}(1-\alpha, \alpha) = \log(1/\alpha)(1 - o(1))$.

Step 2 (Markov correction of the Wald identity). Let $D(c) = D_{\mathrm{KL}}\big(P^{(b)}(\cdot \mid c) \,\|\, P^{(a)}(\cdot \mid c)\big)$ and $\bar{D}_{ba} = \mathbb{E}_{\mu^*}[D]$. Decompose $\Lambda_\tau = \bar{D}_{ba}\tau + \sum_{t<\tau} \big(D(C_t) - \bar{D}_{ba}\big) + \mathcal{M}_\tau$, where $\mathcal{M}_t = \sum_{s<t} \big(\lambda_s - D(C_s)\big)$ is a martingale with bounded increments ($\leq 2\Lambda_{\max}$), and the optional stopping theorem gives $\mathbb{E}_b[\mathcal{M}_\tau] = 0$. For the middle term we solve a Poisson equation. On a finite ergodic chain there exists a bounded h with $(\mathrm{Id} - P_{\text{ctx}}^{(b)})h = D - \bar{D}_{ba}$ (solvability is guaranteed by $\mathbb{E}_{\mu^*}[D - \bar{D}_{ba}] = 0$). In the telescoping identity

$$\sum_{t<\tau} \big(D(C_t) - \bar{D}_{ba}\big) = h(C_0) - h(C_\tau) + \mathcal{M}'_\tau, \qquad \mathcal{M}'_t = \sum_{s<t} \Big[h(C_{s+1}) - \big(P_{\text{ctx}}^{(b)} h\big)(C_s)\Big],$$

$\mathcal{M}'$ is a martingale with bounded increments ($\leq 2\|h\|_\infty$), so $\big|\mathbb{E}_b \sum_{t<\tau} \big(D(C_t) - \bar{D}_{ba}\big)\big| \leq 2\|h\|_\infty$. Combining these gives $\mathbb{E}_b[\Lambda_\tau] \leq \bar{D}_{ba}\, \mathbb{E}_b[\tau] + 2\|h\|_\infty$.

Step 3 (combination). $\bar{D}_{ba}\,\mathbb{E}_b[\tau] \geq \mathrm{kl}(1-\alpha,\alpha) - 2\|h\|_\infty$, that is,

$$\mathbb{E}_b[\tau_{\mathrm{det}}] \;\geq\; \frac{\log(1/\alpha)}{\bar{D}_{ba}}\big(1-o(1)\big) \qquad (\alpha \to 0).$$

□

A.11 Proof of the extension to higher orders (Section 5.2)

Suppose that every strategy is a kernel of order k, all share a strictly positive context stationary distribution μ^*, and the context chain of every strategy is irreducible. With $C_t = (X_{t-k+1}, \ldots, X_t)$ as the context, Lemma 1, Theorem 7 and Proposition 2 hold in parallel.

Order-k version of Lemma 1. The induction of Appendix A.8 proceeds in the same way. In the induction step, $\sum_{c:\, c\to c'} \mu^*(c)P^{(\ell)}(c'_k \mid c) = \mu^*(c')$ is exactly the shared context stationarity. The base case is $C_k \sim \mu^*$ with C_k independent of Z_k.

Order-k version of Theorem 7. The argument of Appendix A.9 proceeds in the same way, taking a context c with $\mu^*(c) > 0$ at which the kernels differ (strict positivity ensures that the event $\{C_t = c\}$ has positive probability).

Order-k version of Proposition 2. The argument of Appendix A.10 proceeds in the same way with $\bar{D}_{ba} = \mathbb{E}_{C\sim\mu^*}\big[D_{\mathrm{KL}}\big(P^{(b)}(\cdot \mid C) \,\|\, P^{(a)}(\cdot \mid C)\big)\big]$, and the solution of the Poisson equation is bounded on the finite ergodic context chain. □

A.12 Proofs of Theorem B.2, Corollary B.2 and Corollary B.2

Theorem B.2. (i) *Class-wise decomposition and dimension.* The constraint $\mu^*(s,j) = \sum_i \mu^*(i,s)P(j \mid i,s)$ involves only the rows $\{(i,s)\}_{i\in\mathcal{X}}$ that share the suffix s, so the constraint set is the product of the constraint sets of the classes. Fix a class s. Summing the n equations ($j \in \mathcal{X}$) over j gives $\sum_j \mu^*(s,j) = \sum_i \mu^*(i,s)$, which is the shift-consistency identity of μ^*, so at most $n-1$ of them are independent. Taking (i_0, s) as the dependent row ($i_0 \in \arg\max_i \mu^*(i,s)$),

$$P(j \mid i_0, s) = \frac{\mu^*(s,j) - \sum_{i\neq i_0} \mu^*(i,s)P(j \mid i,s)}{\mu^*(i_0,s)},$$

whose row sum is automatically 1 (sum over j and use shift consistency). The free dimension of a class is therefore $(n-1)^2$, and the total dimension is $n^{k-1}(n-1)^2$. (ii) *Class-wise factorization of the prior and the likelihood.* The indicator function factorizes as $\mathbf{1}_{\mathcal{C}}(P) = \prod_s \mathbf{1}_{\mathcal{C}_s}(P_s)$, and a product Dirichlet restricted to a product set is still a (restricted) product. In the likelihood $\mu^*(c_k)\prod_t P(x_{t+1} \mid c_t)$, the initial factor does not depend on P (μ^* is fixed), and each transition factor contains only the parameters of the class of its context, so the likelihood factorizes over classes. The posterior is therefore independent across classes, and f^* is computed class by class. (iii) *Weighted first-order problem.* The constraint of class s is $\sum_i \xi_s(i)P(j \mid i,s) = \vartheta_s(j)$, where $\xi_s(i) = \mu^*(i,s)$, $\vartheta_s(j) = \mu^*(s,j)$ and $\sum_j \vartheta_s(j) = \sum_i \xi_s(i)$ (shift consistency). □

Corollary B.2. *(a) $k = 2$, $n = 2$.* Shift consistency makes the marginal of the first coordinate, $\mu_{00} + \mu_{01}$, equal to the marginal of the second coordinate, $\mu_{00} + \mu_{10}$, so $\mu_{01} = \mu_{10}$. Class $s = 0$ has $\xi_s = (\mu_{00}, \mu_{10})$ and $\vartheta_1 = \mu_{01} = \mu_{10} = \xi_s(1)$, and class $s = 1$ has $\xi_s = (\mu_{01}, \mu_{11})$ and $\vartheta_1 = \mu_{11} = \xi_s(1)$. Both classes therefore satisfy $\xi_s = \vartheta_s$ (up to a common normalization). Taking the row with the smaller weight as the free row gives exactly the first-order problem of Section 3 (free parameter p_s, dependent entry $1 - r_s p_s$, $r_s = \min_i \xi_s(i)/\max_i \xi_s(i) \leq 1$, and the whole interval as feasible region), and the ${}_2F_1$ closed form of Appendix A.3 applies class by class. *(b) $k \geq 3$.* Let the weight of the free row be ξ_f and that of the dependent row be ξ_o, with dependent entry $q_s(p_s) = (\vartheta_1 - \xi_f p_s)/\xi_o$. Since q_s is monotone and linear, $q_s([0,1]) \subseteq [0,1]$ if and only if the endpoints satisfy $q_s(0) = \vartheta_1/\xi_o \leq 1$ and $q_s(1) = (\vartheta_1 - \xi_f)/\xi_o \geq 0$, that is, $\xi_f \leq \vartheta_1 \leq \xi_o$. The union over the two choices of the free row gives the necessary and sufficient condition $\vartheta_1 \in [\min_i \xi_s(i), \max_i \xi_s(i)]$. When it holds (in the case $\vartheta_1 < \xi_o$),

$$q_s = \frac{\vartheta_1}{\xi_o}\Big(1 - \frac{\xi_f}{\vartheta_1}p_s\Big), \qquad 1 - q_s = \frac{\xi_o - \vartheta_1}{\xi_o}\Big(1 + \frac{\xi_f}{\xi_o - \vartheta_1}p_s\Big),$$

and the posterior is $\propto p_s^{\widetilde{\kappa}_1-1}(1-p_s)^{\widetilde{\kappa}_0-1}(1-z_1p_s)^{\nu_1}(1+z_2p_s)^{\nu_0}$ ($z_1 = \xi_f/\vartheta_1 \le 1$, $z_2 = \xi_f/(\xi_o-\vartheta_1)$, $\widetilde{\kappa}_1 = \beta + n_{\text{free}\to 1}$, $\widetilde{\kappa}_0 = \beta + n_{\text{free}\to 0}$, $\nu_1 = n_{\text{dep}\to 1}$, $\nu_0 = n_{\text{dep}\to 0}$). A double binomial expansion followed by term-by-term Beta integration gives

$$\mathbb{E}[p_s] = \frac{S(1)}{S(0)}, \qquad S(\iota) = \sum_{j=0}^{\nu_1}\sum_{l=0}^{\nu_0}\binom{\nu_1}{j}(-z_1)^j\binom{\nu_0}{l}z_2^l\, B\big(\widetilde{\kappa}_1+\iota+j+l,\,\widetilde{\kappa}_0\big),$$

which is a ratio of finite double sums (degenerate Appell F_1 polynomials). When $\vartheta_1 = \xi_o$ (that is, $\xi_s = \vartheta_s$), $1-q_s = (\xi_f/\xi_o)p_s$ is a single factor, and the result reduces to the ${}_2F_1$ of Appendix A.3. When the condition fails, the feasible region is truncated to the subinterval $\big[\max\big(0,(\vartheta_1-\xi_o)/\xi_f\big),\ \min\big(1,\vartheta_1/\xi_f\big)\big]$, and the one-dimensional integral of each class can be computed numerically to full precision. □

Corollary B.2. Let μ^* be strictly positive. Take the class $s^\star = 1^{k-1}$, whose two rows are the contexts $(0,1^{k-1})$ (the free row) and 1^k. For $m \ge 0$ define the history $\mathcal{H}_m$ as follows. The initial context is $C_k = 1^k$ (with probability $\mu^*(1^k) > 0$). The symbol 1 is first emitted m times (each time at context 1^k, giving the count $n(1^k \to 1) = m$, and no other context is visited in the meantime), then 0 is emitted once ($n(1^k \to 0) = 1$), and then 1 is emitted $k-1$ times. After this the current context is $(0,1^{k-1})$, and the contexts visited in the final segment ($1^{k-1}0,\ 1^{k-2}01,\ldots$) do not belong to class $s^\star$. The counts of class $s^\star$ are therefore all zero in the free row $(0,1^{k-1})$ and $(m,1)$ in the dependent row 1^k. The history $\mathcal{H}_m$ has strictly positive probability (every transition probability is almost surely positive). $f_\beta(1 \mid (0,1^{k-1})) = \beta/2\beta = 1/2$, which does not depend on m. The posterior of class $s^\star$ (in the free parameter p_s) is multiplied by the factor $q_s(p_s)$ as $m \to m+1$, and this factor is a strictly decreasing affine function (with slope $-\xi_f/\xi_o < 0$). By Lemma A.0, $\mathbb{E}[p_s \mid \mathcal{H}_m]$ is strictly decreasing in m, so f^* takes different values on $\mathcal{H}_m$ and $\mathcal{H}_{m'}$ ($m \ne m'$), and at least one of them is $\ne 1/2 = f_\beta$. By the class-wise independence of Theorem B.2, differences in the counts of other classes do not affect the posterior of this class. Combining this with Lemma A.0 gives the strict suboptimality of order-k add-β. □

B Supplementary theoretical results and numerical verification

B.1 General number of states $n \ge 3$

Take any one row as the dependent row and the remaining $n-1$ rows as free rows. The dependent row is determined by the constraint $\pi^*P = \pi^*$, and its row sum is automatically 1, so the constraint set $\mathcal{C}(\pi^*)$ is an $(n-1)^2$-dimensional polytope with the free rows as coordinates. The entries of the dependent row must be nonnegative, so the free rows are restricted to a feasible region Ω, which for $n \ge 3$ is a proper subset of the product of $n-1$ simplices. The prior is the product $\mathrm{Dir}(\beta)$ on the free rows restricted to Ω. For $n = 2$, the row of the state with the larger stationary probability is taken as the dependent row, and Ω is the whole interval, which gives $p \sim \mathrm{Beta}(\beta,\beta)$ and $q = rp$ of Section 3.

For $n = 2$, Theorem 2 gives the ${}_2F_1$ closed form of f^*. For $n \ge 3$, every standard hypergeometric closed form (of Lauricella or Carlson-R type) is equivalent to removing the truncation to Ω and integrating over the whole product of simplices. Such candidate expressions therefore generally give "probabilities" outside $[0,1]$ and are not valid predictions.

For $n \ge 3$, f^* is an incomplete Dirichlet integral (a Carlson incomplete R function) and must be computed numerically. This conclusion applies only to the standard derivation route via polynomial expansion and Dirichlet moments.

As the number of round trips R grows, the posterior concentrates on a lower-dimensional subset of the feasible region and importance sampling degenerates accordingly. The limit $\mathbb{E}[P_{ij} \mid \text{counts}] \to 0$ in Appendix A.2 is therefore established by Corollary A.2 and Lemma A.2.

Numerical verification (Figure 2(b)). For each $n \in \{2,3,4,5\}$ we take 20 distributions π^*, each being the stationary distribution of a random base kernel whose rows follow $\mathrm{Dir}(5\cdot\mathbf{1})$. The states are labeled in decreasing order of π^*, row 0 is the dependent row, and $\beta = 1$. f^* is computed by

count-adaptive importance sampling, with the free rows' $\mathrm{Dir}(\mathbf{1} + \text{row counts})$ conditioned on Ω as proposal distribution. The likelihood of the free rows cancels with the proposal, and the weights contain only the likelihood factor of the dependent row. When the same estimator is applied to the unconstrained product $\mathrm{Dir}(\mathbf{1})$, f^* equals add-β, and the measured gap is zero within numerical error. Each π^* uses 5 seeds with 200 sequences per seed, the per-step gap is averaged over checkpoints every 3 steps, and $T = 15, 30, 60$ are reported.

For all 80 distributions π^* and all three values of T, the per-step gap is positive (Figure 2(b)). Since f^* is Bayes optimal, its estimation noise can only increase the loss, so the measured gap is biased downward.

B.2 Suffix-class decomposition for $k \geq 2$

In the order-k setting, the context is $c = (i_1, \ldots, i_k) \in \mathcal{X}^k$. Fix a shift-consistent context stationary distribution μ^*, that is, the marginal of its first $k-1$ coordinates equals the marginal of its last $k-1$ coordinates. The constraint is

$$\mu^*(s, j) = \sum_{i \in \mathcal{X}} \mu^*(i, s)\, P(j \mid i, s) \qquad \forall\, s \in \mathcal{X}^{k-1},\ j \in \mathcal{X}.$$

Theorem B.1 (Suffix-class decomposition)**.** *Let the order be $k \geq 2$, let the context stationary distribution μ^* be strictly positive, and suppose that assumption (A2) holds. Then the following three statements hold.*

(i) The constraint links only the rows $\{(i, s)\}_{i \in \mathcal{X}}$ that share the suffix s. The constraint set therefore decomposes by suffix class into n^{k-1} mutually independent subproblems, with total free dimension $n^{k-1}(n-1)^2$.

(ii) The product Dirichlet prior restricted to the constraint set and the sequence likelihood both factorize over classes. A product measure restricted to a product set is still a product, and the likelihood factor of each transition contains only the parameters of the class of its context. The posterior is therefore independent across classes, and f^* is computed class by class.

(iii) Each class is a weighted generalization of the problem of Section 3. The row weights are $\xi_s(i) = \mu^*(i, s)$, the targets are $\vartheta_s(j) = \mu^*(s, j)$, and shift consistency guarantees $\sum_j \vartheta_s(j) = \sum_i \xi_s(i)$.

Remark. The suffix-class decomposition is proposed in this paper. The gaps for $k = 1$ to 4 are shown in Figure 2(c), and the proof is given in Appendix A.12.

Corollary B.1 (Class-wise closed form)**.** *Under the assumptions of Theorem B.2, for $k = 2$ shift consistency forces $\mu_{01} = \mu_{10}$, and hence every class has exactly $\xi_s = \vartheta_s$. Both classes are isomorphic to the first-order problem of Section 3 and have a ${}_2F_1$ closed form class by class, with class parameters (β, r_s), where $r_s = \min_i \xi_s(i) / \max_i \xi_s(i)$. For $k \geq 3$, in general $\xi_s \neq \vartheta_s$. The feasible region of a class is the whole interval if and only if $\vartheta_s(1) \in \big[\min_i \xi_s(i),\ \max_i \xi_s(i)\big]$, with the dependent row chosen accordingly. When this holds, the posterior of each class contains at most two linear non-Beta factors with nonnegative integer exponents, and the predictor is a ratio of finite double sums (degenerate Appell F_1 polynomials), which reduces to ${}_2F_1$ when $\xi_s = \vartheta_s$. When it fails, the feasible region is truncated to a subinterval, which gives an incomplete integral of the same type as in Appendix B.1, but only one-dimensional, so it can be computed numerically to full precision. The proof is given in Appendix A.12.*

Corollary B.2 (Strict suboptimality of order-k add-β for $n = 2$)**.** *Under the assumptions of Theorem B.2, if $n = 2$, then order-k add-β is strictly suboptimal for every $k \geq 2$. Cross-row posterior coupling occurs only within the same suffix class, that is, the counts from row (i', s) change the estimate of row (i, s). The proof is given in Appendix A.12.*

Remark. For $n \geq 3$, the covariance identity within a class gives the same mechanism of cross-row posterior coupling, but strictness requires verifying for the transition graph within each class that closed loops are realizable, and the complete proof is left for future research.

Numerical verification (Figure 2(c)). We take $n = 2$, and for each $k \in \{1, 2, 3, 4\}$ we take 20 distributions μ^*, each being the context stationary distribution of a random order-k base kernel whose transition probabilities follow the uniform distribution on $[0.2, 0.8]$. The free parameter of each class follows a uniform prior on the feasible region of the class ($\beta = 1$), and f^* is computed

class by class by one-dimensional numerical quadrature. Each μ^* uses 5 seeds with 800 sequences per seed, and $T = 30, 60, 120$ are reported.

For all 80 distributions μ^* and all three values of T, the per-step gap is positive. At each T, the mean gap increases with k, and its decay with T becomes slower as k increases, which is consistent with context counts becoming sparser as k increases.

B.3 Asymptotic behavior of the gap and its decomposition under switching

Theorem 1 shows that the gap is strictly positive. This subsection discusses how the gap varies with the number of steps t, the nonuniformity of π^* and the switching rate.

(1) Under a single strategy the gap vanishes at rate $O(1/t)$. Both predictors converge in probability to the same true value p, and they differ in how much information about p they extract at each step. Let J_0 and J_1 denote the Fisher information about p of one transition from state 0 and from state 1, respectively. To estimate the transition probability of row 0, add-β uses only the transitions from state 0, and its Fisher information accumulates at J_0 per step. f^* uses both rows, because the constraint $q = rp$ makes the transitions from state 1 also carry information about p, and its Fisher information accumulates at $J_0 + J_1$ per step. The Bayes excess risk is of the same order as the variance of the parameter estimate, and both are $O(1/t)$, so their difference also tends to zero at rate $O(1/t)$. The coefficient of the leading term is given by Proposition 1 and holds for any finite number of states.

Proposition 1 characterizes the leading order of the cumulative gap for large T. At fixed t, the size of the gap is determined by the nonuniformity of π^*, which is consistent with the information mechanism above. Suppose $\pi_0 \ll \pi_1$. Then state 0 is rarely visited, the count $n_{0\cdot}$ of transitions from state 0 is small, and the estimate of row 0 by add-β has a large error because of the scarce samples. f^* compensates through cross-row posterior coupling. It uses the many visits to state 1 to obtain information about q, which the constraint $q = rp$ converts into information about p. Hence the rarer the state, the larger the benefit of using the constraint (Figure 2(a)).

Numerically, the product of the gap at step t and t approaches the leading coefficient $(n-1)/2$ (Figure 2(d)).

(2) Under strategy switching the gap no longer vanishes. After each strategy switch, the counts accumulated so far no longer apply to the new strategy, so both predictors remain short of samples for long periods, and the $O(1/t)$ decay of (1) has not fully taken effect before the next switch. The total gap can be decomposed into two components, and this decomposition is proposed in this paper.

Structural component. Theorem 3 gives g_{struct} and its lower bound $\log n - H(\pi^*)$ for any n. The lower bound depends only on the known π^* and not on the prior.

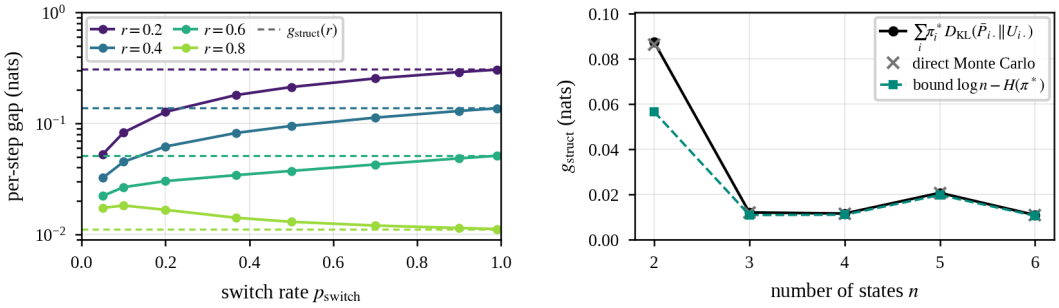
Posterior coupling component. Within a dwell interval the counts accumulate over time and the mechanism of (1) takes effect, so this component decays with the reciprocal of the number of steps elapsed within the interval.

The contributions of the two components to the total gap are as follows. The fraction of steps with zero counts among all steps is exactly p_{switch}, so the contribution of the structural component to the per-step gap is $p_{\text{switch}}\, g_{\text{struct}}$, which increases linearly with the switching rate. The posterior coupling component decays within a dwell interval with the reciprocal of the number of elapsed steps, and its per-step average first increases and then decreases as the intervals become shorter. As $p_{\text{switch}} \to 0$, the dwell intervals become long and both components tend to zero. As $p_{\text{switch}} \to 1$, every step is the beginning of a new interval, only the structural component remains, and the total gap tends to g_{struct} (Theorem 3), and this limit holds for any n. The shape in the intermediate range depends on the relative sizes of the two components. The total gap is a weighted average of the gap g_{struct} at zero counts and the gaps at the subsequent steps within a dwell interval, and the larger p_{switch} is, the larger the weight of the former. Therefore, when g_{struct} is higher than the gaps at the subsequent steps within an interval, the total gap increases with p_{switch}, and when the gap after several steps within an interval is higher than g_{struct}, the total gap decreases with p_{switch}. Figure B.1(a) shows the total gap as a function of p_{switch} for $n = 2$, and the curves for $r = 0.2$ to 0.6 belong to the former case. At

$r = 0.8$, π^* is close to uniform and g_{struct} is small, lower than the gaps after several steps within an interval, so the curve decreases with p_{switch}. The monotonicity of the total gap in p_{switch} therefore depends on π^* and is not a general conclusion.

Remark. For two states, flow balance gives $|n_{01} - n_{10}| \leq 1$, which entangles the two effects of the structural component and the posterior coupling component. An exact decomposition of the two requires $n \geq 3$ and is left for future work.

Numerical values of the structural component are shown in Figure B.1(b). For $n = 3$ to 6, the analytical value of g_{struct} agrees with direct Monte Carlo simulation, and both are above the lower bound of Theorem 3. The simulation for $n = 3$ agrees with this limit as $p_{\text{switch}} \to 1$, and the case $n = 2$ is shown in Figure B.1(a).



(a) $n = 2$, gap versus p_{switch} (b) structural component versus n

Figure B.1: Estimation gap under hidden switching. (a) Per-step gap for two states as a function of the switching rate, where the dashed lines are $g_{\text{struct}}(r)$. (b) The structural component and its lower bound as functions of the number of states.

B.4 Setting and computation method of E1

E1 compares the per-step gap between add-β and f^* under a single strategy as a function of the stationary ratio r and the sequence length T ($n = 2$, $\beta = 1$, Figure 2(a)).

Each data point is a Monte Carlo average over $M = 6000$ random sequences, and f^* is obtained by numerical quadrature of the posterior density of Theorem 2 in the log domain. For $T \gtrsim 160$, the hyp2f1 function of scipy loses precision because of catastrophic cancellation of alternating terms. This paper therefore always computes f^* by numerical quadrature in the log domain rather than by evaluating the hypergeometric function directly.

B.5 Supplementary discussion and numerical results on the statistical limit

Mechanism. The inequality $\rho < 1$ comes from the detection delay under hidden switching and from the uncertainty in predicting switches, that is, the diffusion of the belief by the switching matrix A in the prediction step of the HMM. This is a statistical limit and also holds with infinite memory. Bounded memory is not a separate source. Distinguishing K strategies requires only $\log_2 K$ bits, which is negligible relative to the capacity of order 2^k needed to store context counts, and the limitation of memory reduces to the capacity and precision lower bound of Theorem 2 of Bondaschi et al., which can be removed by increasing the width and the numerical precision. When switches are observable, setting the decay coefficient to zero at a switch to reset the counts is optimal (Bondaschi et al., 2026). When switches are unobservable there is no such trigger signal, and the current strategy must be inferred by the optimal filter.

Numerical results (Figure 5). The value of ρ in each setting is estimated from 5 seeds with 5000 sequences of length 120 per seed. In (a), $K = 2$ and the two strategies take $p \in \{0.3, 0.7\}$. In (b), the values of p of the K strategies are equally spaced on $[0.15, 0.6]$. In (c), $K = 2$, one strategy takes $p = 0.2$, and the p of the other strategy is varied to adjust $\bar{D}_{ba}$. Panels (b) and (c) take $r = 0.6$. For $K = 2$ and $p_{\text{switch}} = 0.5$, the strategy sequence has no temporal correlation, and ρ is zero within estimation noise (Figure 5(a)).

Dependence on the switching rate. The residual loss $\mathcal{L}^* - H_{\text{within}} = (1-\rho)H_{\text{between}}$ can be written as the product of p_{switch} and the loss accumulated per switch, $(1-\rho)H_{\text{between}}/p_{\text{switch}}$. The latter increases as p_{switch} decreases and saturates, which is consistent with Proposition 2. Identifying the new strategy after a switch takes about $\log(1/\alpha)/\bar{D}_{ba}$ steps. When the dwell intervals are much longer than this delay, the loss accumulated per switch tends to a constant determined by $\bar{D}_{ba}$. When the dwell intervals are shorter than this delay, the next switch occurs before the previous one has been identified, and the accumulated loss shrinks accordingly.

Dependence on the number of strategies. ρ is less than 1 for every K, and it varies less with K than with p_{switch} (Figure 5(b)). As the strategies become denser, H_{between} and the residual loss $(1-\rho)H_{\text{between}}$ shrink together, so the ratio ρ changes little, which is consistent with the decomposition above.

Across numbers of states and orders (Figure 6(a)(b)). Figure 6(a) reuses the distributions π^* of Appendix B.1, and for each π^* two strategies are drawn from the prior on the feasible region, subject to $\sum_i \pi_i^* \sum_j |P_{ij}^{(1)} - P_{ij}^{(2)}| \geq 0.4$ and $H_{\text{between}} \geq 0.02$. Figure 6(b) reuses the distributions μ^* of Appendix B.2, and the parameters of the two strategies are taken at the 1/4 and 3/4 quantiles of the feasible region of each suffix class. Both use the optimal filter with $K = 2$, and each π^* or μ^* uses 5 seeds with 200 sequences of length 3000 per seed, with the first 200 steps excluded.

For all π^*, μ^* and switching rates, $\rho < 1$, and the variation of ρ with the number of states and the order is much smaller than its variation with the switching rate. The empirical loss of the oracle predictor agrees with H_{within}, which is a numerical check of Lemma 1.

B.6 Sufficiency of cross-row readouts

Proposition B.1 (Sufficiency of cross-row readouts)**.** *Suppose that the hidden state of an architecture maintains the four transition counts linearly and that its readout is a two-layer feedforward network defined on the count vector. Then for every $\varepsilon > 0$, there exist readout parameters for which the maximum pointwise deviation between the prediction of the architecture and f^* on $\{0, \ldots, T\}^4$ is less than ε.*

Remark. Proposition B.6 turns "the difference lies in the readout" from an empirical observation into a formal statement. The premise that the hidden state maintains the counts linearly is given by the construction of Theorem 1 of Bondaschi et al., and the three trained Mamba architectures also satisfy it (Probe 2, Appendix Figure C.2). The proof is given in Appendix A.6.

C Supplementary experiments

C.1 Supplementary results for E2

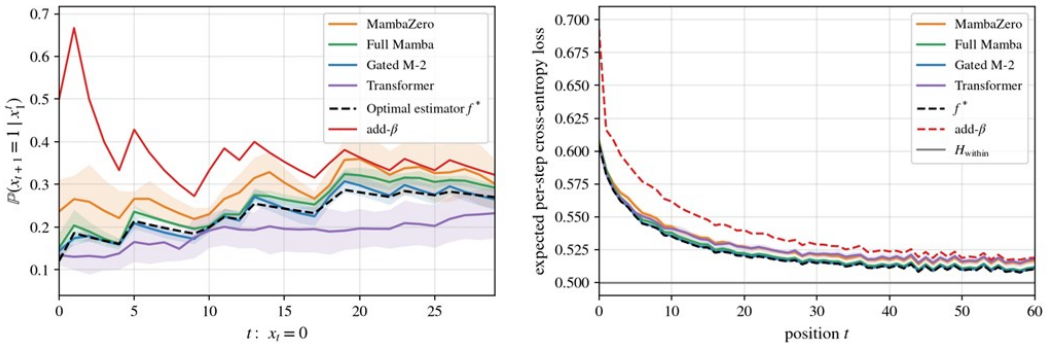
Training dynamics (Figure 3(b)). The closed fraction of the MambaZero variant stays essentially unchanged after 3000 steps, and the loss of the Transformer is still above that of add-β at 4000 steps and continues to decrease afterwards.

Prediction trajectories and per-position loss (Figure C.1). Figure C.1(a) shows the per-step predictions of the four architectures on the held-out sequence used in Figure 3(c), and Figure C.1(b) shows the per-position expected loss on 5000 held-out sequences. The per-position loss of all four architectures is below that of add-β at every plotted position.

C.2 Mechanism probes

The probes in this subsection show that the transition counts are linearly readable from the hidden state, that the counts of all rows are combined at the readout, and that the mechanism for $n = 3$ is the same as for $n = 2$.

Probe 1 (behavioral level). We construct a family of sequences in which the counts of the current row are fixed ($n_{00} = 6$, $n_{01} = 3$, $x_t = 0$, so add-β is constant) and only the count of the other row,



(a) Next-token probability on a held-out sequence ($p = 0.26$)

(b) Expected per-step loss on 5000 held-out sequences

Figure C.1: Prediction trajectories and per-position loss of the models trained in E2 on held-out sequences.

$n_{11} = 3m$ ($m = 0, \ldots, 8$), is varied, which extends the run of consecutive self-transitions of state 1. The predictions of trained full Mamba are shown in Figure C.5(b).

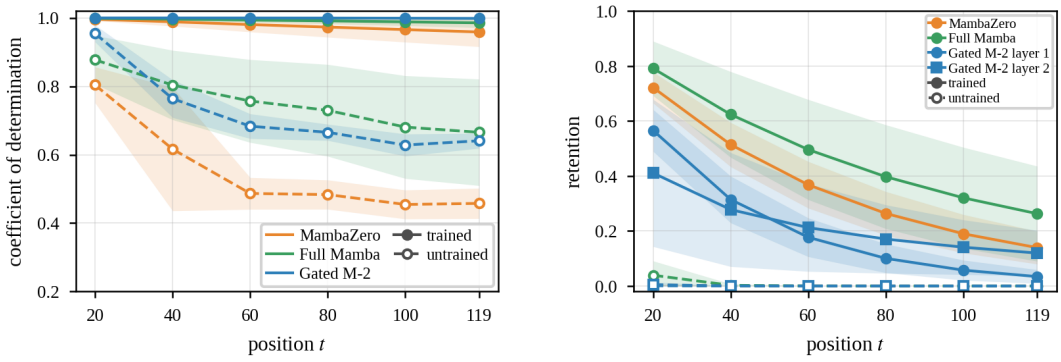
The prediction of the model moves monotonically with n_{11}, with direction and magnitude close to those of f^*, whereas add-β stays unchanged. This is behavioral evidence of cross-row posterior coupling. The constraint makes the $1 \to 1$ transitions also carry information about p, and the model uses this information. In a binary chain the total number of steps is determined by the counts, so changing n_{11} necessarily also changes the sequence length. Since f^* is conditioned on the same counts, the comparison with it is unaffected, but Probe 1 cannot distinguish whether the model uses the counts of the other row or the position. Probe 5 makes this distinction at a fixed length.

Probe 2 (internal linear decoding). On the models trained for 8000 steps in E2, we decode the four transition counts $n_{ij}(t)$ from the hidden state by ridge regression, fitting on the first 4000 held-out sequences and evaluating on the last 1000. The controls are untrained models with the same initialization and regression on shuffled targets (Figure C.2). After training, the coefficient of determination of the three Mamba architectures is no lower than 0.95 from $t = 20$ to $t = 119$. For untrained models it decreases with position and reaches only 0.46 to 0.67 at $t = 119$, and with shuffled targets it is close to zero. The difference comes from the decay coefficient. Training raises the decay coefficient above 0.93. Without training the decay coefficient is at most 0.89, and the first counted transition has almost completely decayed before $t = 40$ (Figure C.2(b)). The coefficient of determination on the input of the readout is lower than on the state, so the counts are further processed in the readout. The Transformer has no recurrent state, and the coefficient of determination on its last-layer representation is lower than on the states of the three Mamba architectures.

Probe 3 ($n = 3$, zero-count predictions and Theorem 3). In the $n = 3$ setting of Appendix C.5 ($\pi^* = (0.20, 0.35, 0.45)$), we train full Mamba ($d = 32$, 10000 training steps), feed only the initial state $x_1 = i$, read its predicted row for x_2, and compare it with the prior mean row $\bar{P}_{i\cdot}$ and the uniform row $U_{i\cdot}$, which are the predictions of f^* and add-β at zero counts, respectively (Figure C.3). The same models are also used for Probe 5 (Figure 4(b)).

At zero counts the prediction of the model is closer to the prior mean row $\bar{P}_{i\cdot}$ than to the uniform row, so the model has encoded the prior knowledge implied by π^* in its parameters (Figure C.3). The π^*-weighted quantity $\sum_i \pi_i^* D_{\mathrm{KL}}(\text{model}_i \,\|\, U_{i\cdot})$ differs from g_{struct} of Theorem 3 by less than 2%.

Probe 5 (test of cross-row information for three states). We construct sequences of length 12 in which the counts of the current row are fixed with only $n_{12} = 1$, and the remaining transitions are split between $n_{22} = m$ and $n_{00} = 8 - m$, so that $f_{\mathcal{F}_2}$ gives the same output for all m. The prediction of full Mamba trained on $n = 3$ varies with m in the same direction as f^*, whereas



(a) linear decoding of the counts

(b) retention of the first counted transition

Figure C.2: Linear decoding of the transition counts from the hidden state (Probe 2). (a) Coefficient of determination of the four counts as a function of position. (b) Retention coefficient of the first counted transition.

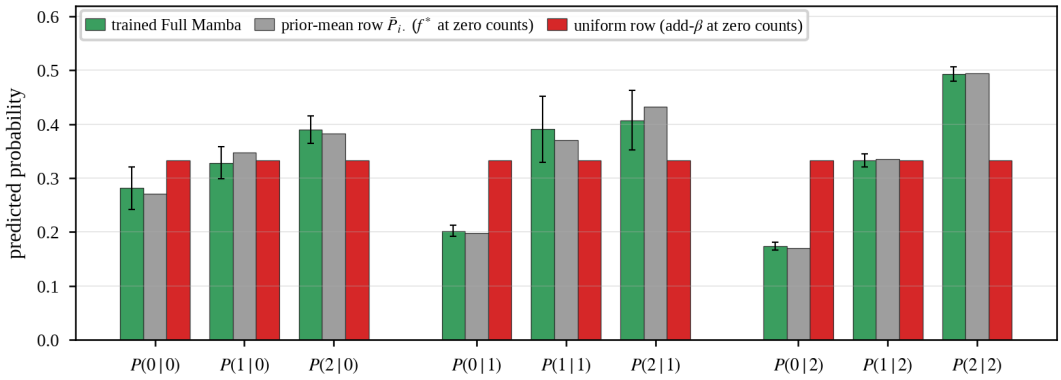
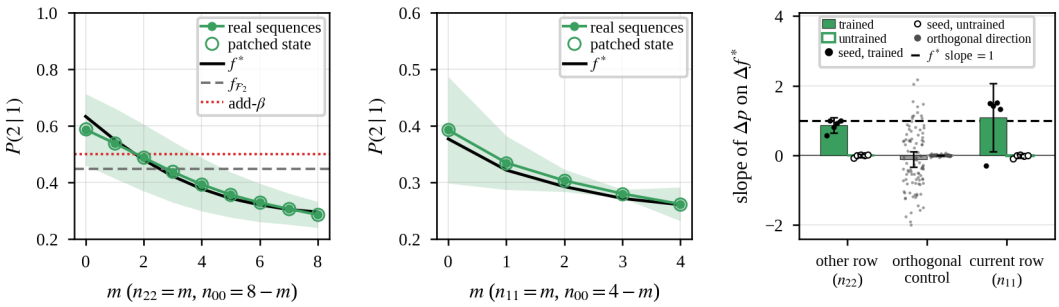


Figure C.3: Zero-count predictions of the trained model for three states (Probe 3).

$f_{\mathcal{F}_2}$ stays unchanged (Figure 4(b) and Figure C.5(c)). The trained model therefore uses cross-row information.

Probe 4 (state patching of cross-row counts). State patching means changing, at one position, only the components of the hidden state that encode a given count while keeping the other components unchanged, and then comparing the predictions of the model. This probe reuses the model and the sequence family of Probe 5 and patches the recurrent state of full Mamba at position $t = 11$. The encoding directions are given by a linear regression of the state on the counts over another set of random sequences, and the fit uses only sequences whose last two symbols are the same as in this sequence family. When the position and the last two symbols are fixed, the state is almost entirely a linear function of the counts. When the state is patched along the encoding direction corresponding to increasing n_{22} and decreasing n_{00} accordingly, the prediction of the model almost coincides with its prediction on the real sequences, and its change is in the same direction as that of f^*, with a magnitude of about 0.86 times the change in f^* (95% interval 0.64 to 1.08), whereas untrained models show no such response (Figure C.4). Patching the self-loop of the current row serves as a positive control, and its response is in the same direction as f^* on most seeds, but it varies considerably across seeds. Perturbations of the same norm orthogonal to the count-encoding directions have an effect on the prediction of comparable size, but without a consistent direction relative to the change in f^*, and the slopes of the individual directions scatter around 0. At this position and with these last two symbols, the state has almost no variance along these directions, so this response does not indicate that the readout uses information other than the counts.



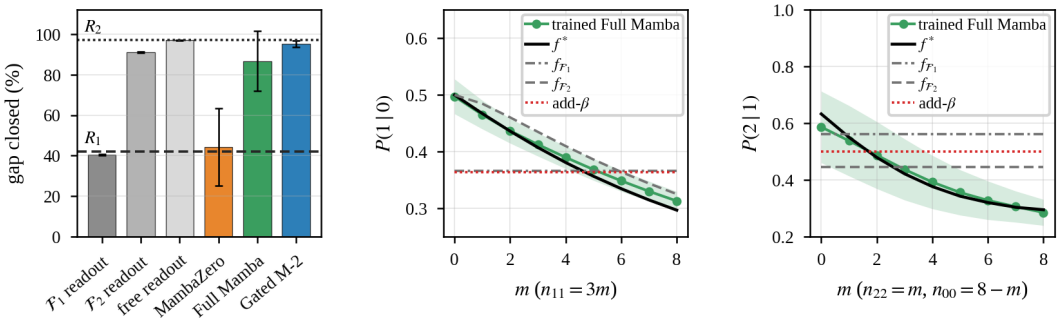
(a) other-row patch (b) current-row self-loop patch (c) response slopes

Figure C.4: State patching of cross-row counts (Probe 4). (a) Patching of other rows, starting from the sequence with $m = 4$. (b) Patching of the self-loop of the current row, as a positive control. (c) Slope of the change in prediction against the change in f^*. All intervals and error bars are 95% intervals over 5 seeds. In (c), filled bars are trained models, hollow bars are untrained models, dots are the values of individual seeds, and gray dots are the values of individual orthogonal directions. The bar heights of the orthogonal control are averaged first over directions and then over seeds.

C.3 Readout ablation

Upper bounds for within-row readouts and readout ablation ($n = 2$, E2 setting, Figure C.5(a)). The predictors $f_{\mathcal{F}_1}$ and $f_{\mathcal{F}_2}$ of Theorem 4 are estimated by averaging f^* within information bins over a large-scale simulation. To test whether these upper bounds can be reached by learned readouts, we train three readout models with the exact counts as input, all two-layer feedforward networks of width 64.

There are three conclusions. First, within-row readouts are strictly bounded by their upper bounds, and the learned within-row readout reaches the upper bound. Second, for $n = 2$, adding the position brings the upper bound close to f^* (by flow balance, see Section 4.1), so for $n = 2$ the closed fraction cannot distinguish a cross-row readout from a within-row readout with position. Third, the closed fraction of the MambaZero variant lies between the two upper bounds, which is consistent with the capability of its normalized readout.



(a) readouts and trained models, $n=2$ (b) probe 1, $n=2$ (c) probe 5, $n=3$

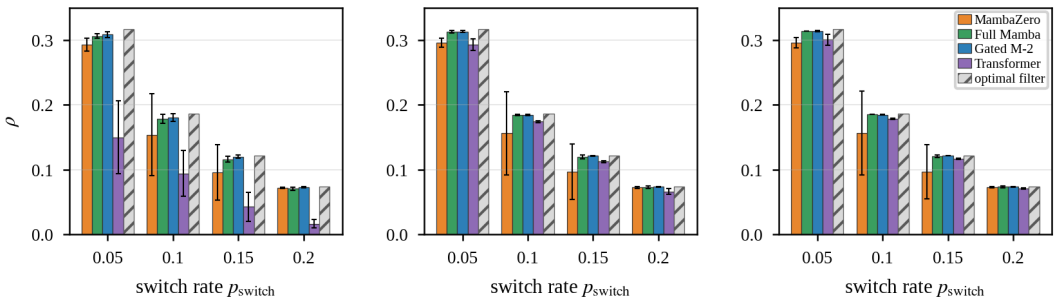
Figure C.5: Ablation of the readout structure and the capability gap. (a) Upper bounds for within-row readouts, learned readouts, and the closed fractions of the three Mamba architectures for two states. (b) Probe 1, in which only the counts of the other row are varied for two states and the sequence length changes accordingly. (c) Probe 5, in which the length and the counts of the current row are fixed for three states and only the counts of the other rows are varied. All intervals and error bars are 95% intervals over 5 seeds.

C.4 Supplementary results for E3

For all switching rates and training budgets, every seed of every architecture satisfies $H_{\text{within}} < \mathcal{L}^* \leq \mathcal{L}_{\text{model}} < H_{\pi^*}(\bar{P}^w)$, that is, the degrees of identification of the trained models and of the optimal

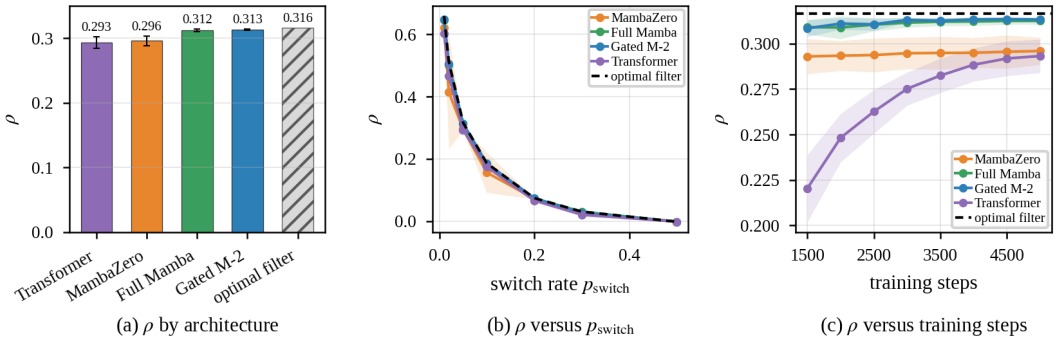
filter satisfy $\rho^{\text{model}} \leq \rho^{\text{HMM}} < 1$ (Figures C.6 and C.7). The variance across seeds in E3 is much smaller than in E2. Tracking the hidden strategy is easy to learn to near optimality, while the optimal filter itself is limited by $\rho < 1$.

Readout-off positions. A position at which all preactivations of the readout layer are nonpositive is called a readout-off position. On switching data of order $k = 1$ to 4 ($p_{\text{switch}} = 0.05$, five random μ^* per order, Figure 6(d)), the readout of Gated M-2 remains essentially active after training, and the median fraction of readout-off positions is at most 4%. The degree of identification of the MambaZero variant falls behind the optimal filter from $k = 3$ on, and at $k = 4$ the median fraction of its readout-off positions rises to about 30%.



(a) 1500 training steps (b) 5000 training steps (c) 8000 training steps

Figure C.6: Degree of identification of the four architectures in E3 under three training budgets.



(a) ρ by architecture (b) ρ versus p_{switch} (c) ρ versus training steps

Figure C.7: Degree of identification of the models trained in E3 as a function of the architecture, the switching rate and the number of training steps. (a) $p_{\text{switch}} = 0.05$, 5000 training steps. (b) 5000 training steps. (c) $p_{\text{switch}} = 0.05$.

C.5 Extensions to more states and higher orders

Three-state setting. $\pi^* = (0.20, 0.35, 0.45)$, row 0 is the dependent row, and the prior is $(P_{1\cdot}, P_{2\cdot}) \sim \text{Dir}(1)^2$ restricted by rejection sampling to the feasible region $\Omega = \{P_{0\cdot} \geq 0\}$. f^* has no closed form and is computed with the importance-sampling estimator of Appendix B.1.

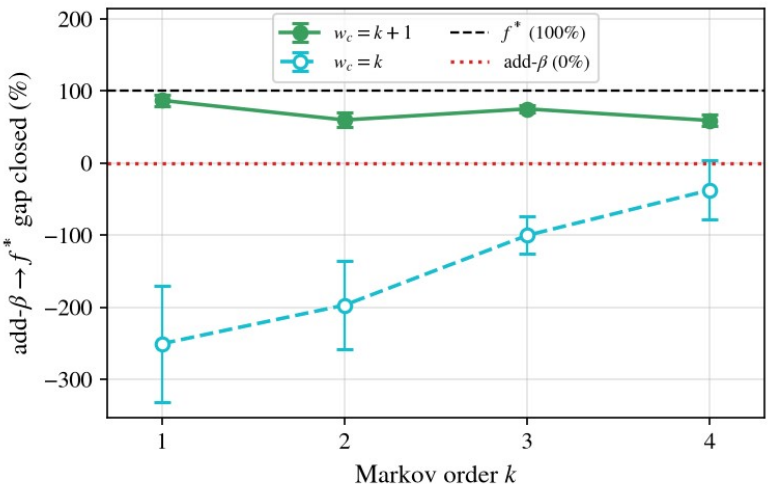
E2-n3 (three states). We train full Mamba ($d = 32$, 10000 steps, $T = 120$) on single-strategy data of this setting and evaluate it every 3 steps on 500 held-out sequences. The gap between add-β and f^* is larger than for $n = 2$. For $n = 2$, f^* has only one free parameter, so the expectation of limited capacity in Section 3.4 does not hold. For $n = 3$, f^* is a posterior mean over a 4-dimensional feasible region, and the closing after training is partial.

E2-orders (orders $k = 1$ to 4). We train full Mamba ($d = 32$, 15000 steps) on order-k single-strategy data, with 5 random μ^* per order. The convolution window length w_c takes the two values $k+1$ and k, and the former satisfies the window condition $w_c \geq k+1$ under which Bondaschi et al. represent order-k add-β (Figure C.8).

With $w_c = k + 1$, the mean closed fraction is positive at every order, and for $k \geq 2$ it is lower than for $k = 1$, which is consistent with the results across numbers of states (Figure C.9). With $w_c = k$, the mean loss at every order is above that of add-β.

E2-states (numbers of states $n = 2$ to 5). We use the n-ary generalization of full Mamba (n-ary embedding table and n-way softmax readout, $w_c = 2$), with 5 random π^* per n and widths $d = 16$ and $d = 32$ (Figure C.9).

With $d = 16$, the mean loss for $n \geq 3$ is above that of add-β under both numbers of training steps, and doubling the number of training steps only narrows this gap without eliminating it. After doubling the width, the mean closed fractions for $n = 3$ and $n = 4$ become positive, and for $n = 5$ it is only slightly above zero. Hence for $n \geq 3$ the width is the dominant factor. This is consistent with Section 4.1, since for $n \geq 3$ the position cannot substitute for cross-row information, and the readout layer needs capacity to represent the cross-row structure.



(a) Orders: full architecture, $w_c = k + 1$ vs $w_c = k$

Figure C.8: Closed fraction after training across orders and convolution windows.

C.6 Recurrence under a finite strategy set

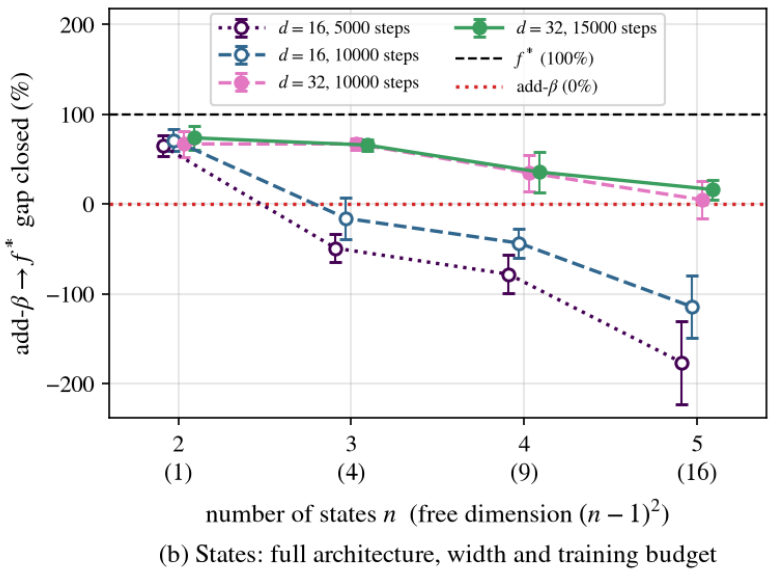
When the strategy set $\mathcal{P}$ is finite and each strategy is selected repeatedly, a predictor can in principle classify strategies that have already appeared instead of re-estimating each time. In experiment E3, full Mamba and Gated M-2 track the finite strategy set nearly optimally (Figure 6(c)).

Relative to a setting in which every dwell interval uses a fresh transition kernel, the advantage brought by strategy recurrence is small for $n = 2$, and its sign depends on the switching rate and the evaluation protocol. The evaluation assumes known switching times and resets the counts at each switch, and this protocol is not symmetric between recurrence and non-recurrence. A two-state transition kernel has only one free parameter, so the cost of re-estimation is already low. This cost rises as n grows, and the advantage brought by recurrence is expected to grow accordingly.

For trained Mamba, the residual loss remains strictly positive under strategy recurrence, and this holds for any number of strategies (Section 5). How large the advantage of recurrence is for $n \geq 3$ is left for future work.

C.7 Design objective in the adversarial setting (estimation gap)

Sections 3 and 4 characterize the gap between add-β and f^* from the perspective of the predictor. This subsection writes this gap as a design objective from the perspective of the patroller, that is, the defender. The defender chooses the strategy set $\{P^{(\ell)}\}$ and the switching rate p_{switch} to enlarge the



(b) States: full architecture, width and training budget

Figure C.9: Closed fraction after training across numbers of states, widths and numbers of training steps.

gap, that is,

$$\max_{\{P^{(\ell)}\},\, p_{\text{switch}}} \mathbb{E}\big[D_{\text{KL}}(f^* \,\|\, f_\beta)\big] \quad \text{s.t. (i)–(iv)},$$

where the constraints (i) to (iv) are, in order, the topology constraint, nonnegativity, row stochasticity and the stationarity constraint.

This objective is effective only against predictors of the add-β type. Trained Mamba closes most of the estimation gap (Section 3.4), so enlarging the estimation gap does not constitute a robust defense. For trained sequence models, the controllable quantities are given in Appendix C.8.

C.8 Controllable quantities in the adversarial setting

For any predictor f, $\mathcal{L}(f) \geq \mathcal{L}^* = H_{\text{within}} + (1 - \rho^{\text{HMM}})H_{\text{between}} > H_{\text{within}}$. Proposition 2 shows that the number of steps needed to identify the new strategy after a switch is at least proportional to $1/\bar{D}_{ba}$, so lowering the strategy separability $\bar{D}_{ba}$ lengthens detection and lowers ρ. The similarity of the same set of strategies determines two quantities at once. H_{between} is an upper bound on the residual loss, and $\bar{D}_{ba}$ determines the speed of identification, so the design of the defender is a trade-off between the two.

In Figure 5(c), ρ increases with $\bar{D}_{ba}$ at every switching rate (the setting is given in Appendix B.5), which is consistent with the detection-delay mechanism of Proposition 2. The more separable the strategies, the faster the identification. H_{between} also increases with separability, and the residual loss $(1-\rho)H_{\text{between}}$ is determined by both. When the separability is very low, ρ is zero within estimation noise, but H_{between} itself is then small. By Figure 5(b), the main controllable quantity is the degree of difference between strategies rather than the number of strategies.

D Experimental details

D.1 Protocol and architectures

Experiment naming. E1 to E3 are the experiment numbers of this paper. The suffixes -n3 and -states refer to settings over the number of states, and -orders refers to settings over the order.

Tested architectures. The MambaZero variant consists of a convolution with window 2, a selective recurrence, a residual connection and an ℓ_1-normalized readout with learnable additive smoothing. The gated simplified architecture Gated M-2 is a two-layer simplified Mamba-2 with gated SiLU and residual connections. Full Mamba retains all components of the original Mamba. The Transformer is a two-layer GPT-2-style model with learned positional embeddings and single-head causal attention.

Number of seeds. Consistent with the convention of Appendix F of Bondaschi et al. (2026), all training experiments use 5 random seeds. Numerical computations such as closed forms, filtering and quadrature involve no training, and their accuracy is determined by Monte Carlo sampling error or numerical grid error.

D.2 Training settings

Data of E2. Constrained single-strategy Markov sequences with $n = 2$, $r = 0.5$, $\beta = 1$ and $q = rp$, where each sequence draws one $p \sim \mathrm{Unif}(0, 1)$.

Readout of the MambaZero variant. The prediction is

$$\widehat{f}(1) = \frac{\mathrm{relu}(v_1) + \widehat{\beta}_1}{\big(\mathrm{relu}(v_0) + \widehat{\beta}_0\big) + \big(\mathrm{relu}(v_1) + \widehat{\beta}_1\big)},$$

where v_0, v_1 are the two outputs of the linear readout layer before the relu, corresponding to the two states, and $\widehat{\beta}_0, \widehat{\beta}_1 > 0$ are learnable additive smoothing parameters that play the smoothing role of β. This parameterization contains add-β as a special case. Relative to the MambaZero of Bondaschi et al., the changes are only in the smoothing parameters and the normalization, hence the name variant.

Full Mamba. It is reproduced strictly according to the definition in Section 2.2 of Bondaschi et al. (2026), including the three-way convolution, ReLU selectivity, input-dependent decay coefficients, $2\times$ internal expansion, output gating, residual connections, the gated MLP (Appendix B of that paper) and the softmax readout.

Training protocol. The training and evaluation lengths of E2 and E3 are both $T = 120$, each number of training steps is trained from scratch, and the hyperparameters of each architecture are the same in the two sets of experiments. The model width is denoted d. The MambaZero variant uses $d = 8$, Adam with constant learning rate 3×10^{-3} and batch size 128. Gated M-2 uses $d = 8$, two layers, Adam with constant learning rate 2×10^{-3} and batch size 128. Full Mamba uses $d = 8$, SSM state dimension 8, $2\times$ internal expansion, AdamW with momentum coefficients 0.9 and 0.95, learning rate 10^{-3} with cosine decay over the training steps, weight decay 10^{-3} and batch size 64. The Transformer uses $d = 16$ with pre-LN and a $4d$ GELU MLP, the same optimizer as full Mamba and batch size 64. E2 is evaluated on 5000 held-out sequences generated with seed 777, and E3 is evaluated on 5000 sequences generated with seed 555.

Extension experiments. The extension experiments over the number of states and the order (Appendix C.5 and Figure 6(d)) and the three-state probes use model widths larger than 8. For the Mamba-type architectures, after the original initialization, the tensors whose input dimension grows with d are multiplied by $\sqrt{8/d}$, and the other tensors are unchanged. The extension over the number of states uses $d \in \{16, 32\}$, the extension over the order uses $d = 32$ with 15000 training steps, and the three-state probes use $d = 32$ with 10000 training steps.

E Application background

The setting of this paper comes from a structure shared by two fields. In adversarial patrolling by unmanned systems (Alvarenga et al., 2019; Wang et al., 2026), the transition kernels of the defender share a stationary distribution because of coverage requirements and switch over time among several such kernels, and the attacker can only learn from the patrol trajectory to predict the next position. In

industrial wireless communication, the U.S. Federal Communications Commission (Federal Communications Commission, 2025) requires frequency-hopping systems to use each frequency equally on average, an attacker can recover the hopping sequence from observations and carry out selective jamming (Cheng et al., 2020), and the defender randomly permutes the channel usage pattern (Tiloca et al., 2019). For defense design in these fields, the conclusion of this paper is that provable unpredictability must rest on the statistical limit created by hidden switching rather than on the difficulty of estimation under a single strategy.

F TABLE OF SYMBOLS

Table F.1 lists the symbols used throughout the paper. Notation used locally within a proof is defined in that proof and is not listed in the table.

G RELATED WORK

In-context learning on Markov chains. For Transformers, prior work has studied the formation of statistical induction heads, training dynamics, the required depth, the effect of tokenization, and the connection between self-attention and Markov models (Edelman et al., 2024; Nichani et al., 2024; Rajaraman et al., 2024a; Sanford et al., 2024; Ekbote et al., 2025; Makkuva et al., 2024; 2025; Rajaraman et al., 2024b; Ildiz et al., 2024). The object of study has also been extended to variable-order Markov chains and to in-context estimation when the test transition matrix differs from the training matrices (Zhou et al., 2026; Lepage et al., 2025). In these works and in Bondaschi et al. (2026), the transition kernel is either fixed or has independently sampled rows. This paper studies the case in which the rows are coupled through a shared stationary distribution.

Mechanisms of in-context learning and differences between architectures. Induction heads are regarded as the core circuit of in-context learning in Transformers (Olsson et al., 2022; Reddy, 2024). On regular languages, Transformers clearly outperform recurrent and convolutional models including Mamba (Akyürek et al., 2024), while other comparisons reach mixed conclusions (Grazzi et al., 2024; Park et al., 2024; Halloran et al., 2025), and architectures with similar performance may still differ in their internal mechanisms (Wang et al., 2025). On the theoretical side, prior work has shown that standard linear recurrent updates cannot directly implement gradient descent in context, whereas adding a multiplicative readout and a cross-product update over a short window makes this possible (Tian et al., 2026). Mamba can learn low-dimensional targets through test-time feature learning (Oh et al., 2025), and its expressive power has also been analyzed (Muca Cirone et al., 2024). This paper combines probes with representation results on the readout structure to examine how Mamba reads out once the constraint changes the optimal solution.

Hidden switching and sequential detection. In Bondaschi et al. (2026), switches are marked by a special symbol, and Mamba resets the counts by setting the decay coefficient to zero. When switches are invisible, the optimal predictor must infer the current transition kernel by forward filtering. The behavior of sequence models on hidden Markov data and on latent mixture data has been studied (Shai et al., 2024; Hu et al., 2024; D'Angelo & Flammarion, 2026), and detection delay is constrained by sequential detection theory (Wald, 1945; Page, 1954; Lorden, 1971; Lai, 1998). This paper uses the instantaneous independence under a shared stationary constraint (Lemma 1) to prove that the residual loss is strictly positive.

Table F.1: Symbols used throughout the paper.

Symbol	Meaning	Defined in
n	Number of states	§2.1
$\mathcal{X}$	State space $\{0, \dots, n-1\}$	§2.1
$X_t,\ x_t$	State at time t and its observed value	§2.1
$P,\ P_{ij}$	Transition kernel and its entry in row i and column j, also written $P(j \mid i)$	§2.1
$\mathbb{P},\ \mathbb{E}$	Probability and expectation	§2.1
π^*	Target stationary distribution	§2.1
$\mathcal{C}(\pi^*)$	Constraint set, that is, all transition kernels with π^* as invariant distribution	§2.1
$\mathcal{P},\ P^{(\ell)}$	Strategy set and its ℓ-th strategy	§2.1
K	Number of strategies	§2.1
Z_t	Index of the strategy in effect at time t	§2.1
A	Switching matrix, that is, the transition matrix of Z_t	§2.1
w	Stationary distribution of A	§2.1
p_{switch}	Total switching probability	§2.1
$\bar{P}^w$	Strategy-averaged kernel $\sum_\ell w_\ell P^{(\ell)}$	§2.1
$H(\cdot),\ H_{\pi^*}(P)$	Entropy, and the conditional entropy $\sum_i \pi_i^* H(P_{i\cdot})$ of kernel P under π^*	§2.1
$H_{\text{within}},\ H_{\text{between}}$	Within-strategy entropy and between-strategy entropy	§2.1
β	Prior pseudo-count, also the smoothing parameter of add-β	§2.1
f	Predictor	§2.2
x_1^t	Observed history $x_1, \dots, x_t$	§2.2
$\mathcal{L}_t(f),\ \mathcal{L}(f)$	Expected log loss of predictor f at step t, and the per-step loss, that is, the average of the losses over steps	§2.2
$\mathcal{L}_t^*,\ \mathcal{L}^*$	Optimal loss at step t, and the average of the optimal losses over steps	§2.2
$\text{excess}_t(f)$	Excess risk at step t	§2.2
ρ	Degree of identification of the strategy currently in effect	§2.2
n_{ij}	Number of $i \to j$ transitions in the sequence	§2.2
f_β	The add-β predictor, that is, Laplacian smoothing	§2.2
r	Stationary ratio π_0/π_1 for two states	§3.1
$p,\ q$	Two-state transition probabilities $P(1 \mid 0)$ and $P(0 \mid 1)$	§3.1
f^*	Constrained Bayes predictor	§3.2
$\kappa_0,\ \kappa_1$	Count combinations $\beta + n_{00}$ and $\beta + n_{01} + n_{10}$ in the two-state closed form	§3.2
T	Sequence length, also called the horizon	§2.2
$D_{\text{KL}}(\cdot \,\|\, \cdot)$	KL divergence	§3.3
Π	Prior supported on the constraint set	§3.3
$\bar{P}$	Prior mean kernel $\mathbb{E}_\Pi[P]$	§3.3
U	Transition kernel whose rows are all uniform	§3.3
g_{struct}	Structural component at zero counts	§3.3
$\mathcal{F}_1,\ \mathcal{F}_2$	Predictor classes of within-row readouts, with $\mathcal{F}_2$ also taking the position as input	§4.1
$f_{\mathcal{F}_1},\ f_{\mathcal{F}_2}$	Bayes-optimal predictors in $\mathcal{F}_1$ and $\mathcal{F}_2$	§4.1
$\mathbf{n}$	Two-state count vector $(n_{00}, n_{01}, n_{10}, n_{11})$	§4.2
ω	Weight function of the Gauss–Jacobi quadrature	§4.2
$p_g,\ \omega_g$	Nodes and weights of the Gauss–Jacobi quadrature	§4.2
L	Readout width	§4.2
$\mathbf{N}_t$	Transition count matrix	§4.2
e_x	One-hot vector of symbol x	§4.2
$a,\ b$	Strategies before and after a switch	§5.2
$\bar{D}_{ba}$	Strategy separability, that is, the average KL divergence of strategy b from strategy a when the state follows π^*	§5.2
τ_{det}	Detection delay	§5.2
α	Upper bound on the false-alarm and missed-detection probabilities	§5.2
k	Order of the Markov chain	§5.2
μ^*	Order-k context stationary distribution	§5.2
C_t	Context $(X_{t-k+1}, \dots, X_t)$ at time t	§5.2
d	Model width	Appendix D.2
w_c	Convolution window length	Appendix C.5