hexagon:2611.00969v1

Time Reversal and the Memory Cost of Prediction

Zhongpai Gao

Abstract

A predictor must summarize the past by one of L states to forecast the entire future. For a source with two hidden states and four symbols, we show that the optimal excess log loss has order L^{-5} forward and L^{-9} backward. Two symbols reset the hidden-state distribution. In a family where a reset retains a dependence η on the past, the reverse L^{-9} law persists up to budgets of order $\eta^{-1/2}$. Beyond that threshold we determine the reverse exponent for $L = \lfloor \eta^{-A} \rfloor$, $1/2 < A \leq 2$. A memory that updates after each symbol attains these rates. For a positive two-hidden-state source with invertible symbol matrices and dependent observations, word probabilities and determinants give a common upper bound in both directions. It has the correct power order in each direction where different final symbols give disjoint sets of posterior probabilities. Separation in either direction gives equality of the logarithmic exponents. We give conditions under which three-symbol words recover both loss curves and examples where even short words and cycles together fail to determine their exponents.

Index terms: Predictive compression, hidden Markov models, quantization, time reversal.

1 Introduction

A predictor knows the law of a stationary sequence and has observed its entire past. With only L memory states, which histories should share a forecast? We measure the price of grouping them by the mean relative entropy from the future law given the full past to the law given the retained state. Let $d_L^{\rightarrow}$ be its infimum and $d_L^{\leftarrow}$ the corresponding infimum when the sequence is read backward. These minima allow any summary of the past. We state separately when a bound is attainable by updating the state after each symbol. Logarithms are natural, and $a_L \asymp b_L$ means that their ratio stays between positive constants. A loss $L^{-\beta+o(1)}$ specifies its logarithmic exponent.

Our first result shows how much the direction can matter even when the source and its alphabet stay fixed as the memory budget grows. Here a symbol matrix T^a records the probabilities of the hidden transitions that emit a. The scalar closed-walk data of the source are the coefficients of $\det(I - z\sum_a u_a T^a)$ with scalar weights u_a.

Theorem 1.1. *There is a stationary source with two hidden states and four symbols, with all labelled transition probabilities at least* $1/16$*, for which the optimal whole-future losses in the two time directions satisfy*

$$d_L^{\rightarrow} \asymp L^{-5}, \qquad d_L^{\leftarrow} \asymp L^{-9}.$$

Both upper bounds are attained in order by a deterministic state updated at each new symbol of the past. The forward and reverse labelled presentations have identical scalar closed-walk determinants for every choice of symbol weights.

A reset erases the effect of preceding observations. After k continuation symbols there are 2^k possible words, with total probability 2^{-k-1}. For a given reset sign, the closest posterior means of

the hidden sign are separated on scales 4^{-k} forward and 16^{-k} backward. The rates combine these squared distances with the run probabilities. Section 4.2 gives the four matrices.

What changes when every symbol matrix is invertible? With two hidden states the posterior is specified by one probability. After a word w, the length of the interval of possible posteriors scales by a factor comparable to $|\det T_w|/p_w^2$, where p_w is the probability of that word. Weighting its squared length by p_w gives the quantity in the following theorem. Reversing both the source and the word preserves that quantity.

Theorem 1.2. *Let a stationary two-hidden-state source have a finite alphabet, a non-i.i.d. observed law, and strictly positive, invertible symbol matrices T^a. For a word w, write p_w for its stationary probability and T_w for its matrix product in chronological order. The function*

$$P(s) = \lim_{n\to\infty} \frac{1}{n} \log \sum_{|w|=n} \left(\frac{|\det T_w|^2}{p_w^3} \right)^s \tag{1}$$

has a unique zero $s_* \in (0,1)$. Put $\beta = 1/s_* - 1$. Both directions admit deterministic memory updates with excess loss $O(L^{-\beta})$.

In a given direction, let K be the support of the posterior probability of the first hidden state given the entire past, and let F_a be its update when the next observed symbol is a. If the sets $F_a(K)$ are pairwise disjoint in that direction, the optimal excess loss has order $L^{-\beta}$ there. If this condition holds in either direction, both logarithmic exponents $\lim_{L\to\infty}(-\log d_L)/\log L$ equal β.

The logarithmic conclusion allows a factor $L^{o(1)}$. We prove the separated-direction bounds directly by a stopping-word argument, following conformal quantization [18, 25] and checking the source's word probabilities and posterior updates. Only the lower bound in a direction that may overlap uses external dimension results: Rapaport's preprint [24, Theorem 1.2] and Feng–Hu [4, Theorem 2.8], applied to independent blocks in Appendix A.6.1.

Section 5 studies the boundary between these two results. Arbitrarily near the reset example are invertible sources with a common exponent approaching 5. When only the resets change, their error η and the budget L determine how long the original reverse rate L^{-9} persists. We distinguish the limit at fixed η from the limit as η decreases and L grows.

Relation to earlier work. Causal states group histories with identical future laws [10, 11]. Marzen and Crutchfield [20, Theorem 1 and Section 5.3] reduced whole-future predictive compression to these states and compared the two directions for the Random Insertion Process, which has finitely many forecasts each way. Our fixed source has infinitely many in each direction. For hidden Markov sources, Marzen and Crutchfield also bounded the loss by covering the possible hidden-state distributions with cubes [21]. Under positivity, our quadratic loss estimate halves the exponent in their upper bound on the number of cubes required. Theorem 4.1 gives the hypotheses for a matching lower bound. Pressure formulas for separated conformal quantization are classical [25, Theorem 3.1 and Lemma 3.2]. Our separated-direction formula uses this framework. Kesseböhmer–Niemann–Zhu [16, Corollary 1.12] also treat Gibbs measures without separation. Here the additional steps are the one-sided-separation reversal transfer and the reset perturbation analysis. For reading-direction effects in language models, see [23].

Sections 3–4 develop the loss comparison and fixed example. Section 5 studies perturbations, and Section 6 recovers rates from short observations. Appendix B treats next-symbol forecasts and the role of positivity.

2 Prediction loss and finite memory

2.1 Prediction from a retained state

A forecast q of a symbol Y incurs log loss $-\log q(Y)$. Let C contain all information from the past relevant to predicting Y, and let S be the retained state, possibly using independent randomness. Write p_C and p_S for the corresponding conditional laws. Since Y is conditionally independent of S given C, $p_S = \mathbb{E}[p_C \mid S]$. Kullback–Leibler divergence $D_{\mathrm{KL}}(p\|q)$ is the mean of $\log(p(Y)/q(Y))$ under p. The expected excess loss of a forecast q_S splits as

$$\mathbb{E}D_{\mathrm{KL}}(p_C\|q_S) = \mathbb{E}D_{\mathrm{KL}}(p_C\|p_S) + \mathbb{E}D_{\mathrm{KL}}(p_S\|q_S). \tag{2}$$

Split $\log(p_C(Y)/q_S(Y))$ at $p_S(Y)$ and condition on S in the second term. Thus p_S is the best forecast from the state, and the first term measures the loss of retaining only S. Apply the same identity to finite future blocks and take the increasing limit to define loss for the entire future. We use R_L for a next-symbol loss and d_L for a whole-future loss.

For a stationary Markov chain, C is the current symbol. Once the L state forecasts are fixed, assigning each symbol to its cheapest forecast minimizes the loss. Earlier observations and randomization cannot reduce it. An optimal state is therefore a function of the current symbol. It uses $\lceil \log_2 L \rceil$ bits. The source law and prediction rule are given.

Proposition 2.1 (Continuity at a fixed memory budget)**.** *Fix a finite alphabet and $L \geq 1$. Among strictly positive two-hidden-state sources, $d_L^{\rightarrow}$ and $d_L^{\leftarrow}$ are continuous functions of the labelled transition probabilities. Symbol matrices may be singular.*

The proof is in Appendix A.6.1. Thus the jumps below concern asymptotic exponents as $L \to \infty$, rather than a loss at a fixed budget.

2.2 What scalar cycle data preserve

For an observed Markov chain with transition matrix P, give symbol i weight d_i and set $D = \mathrm{diag}(d_1, \ldots, d_n)$. The trace of $(PD)^k$ sums the weights of all closed walks of length k. Each walk contributes the product of its transition probabilities and symbol weights. With $\mathcal{Z}_D(P; z) = \det(I - zPD)^{-1}$, the coefficient of z^k in $\log \mathcal{Z}_D(P; z)$ is $\mathrm{tr}((PD)^k)/k$.

If π is the positive stationary distribution and $\Pi = \mathrm{diag}(\pi_1, \ldots, \pi_n)$, the reverse chain has matrix $P^{\leftarrow} = \Pi^{-1}P^{\mathsf{T}}\Pi$. Since diagonal matrices commute,

$$\det(I - zP^{\leftarrow}D) = \det(I - zP^{\mathsf{T}}D) = \det(I - zDP) = \det(I - zPD). \tag{3}$$

Thus the two directions have identical scalar cycle data. The coefficient of $z^{|J|}\prod_{j\in J} d_j$ in the determinant is $(-1)^{|J|}$ times the principal minor indexed by J [19]. For a hidden chain, reversal sends each symbol matrix T^a to $\Pi^{-1}(T^a)^{\mathsf{T}}\Pi$, with Π formed from the stationary hidden law. The determinant with separate scalar weights for the symbols is again preserved. The next section compares the prediction losses of sources with precisely this ambiguity.

3 An observed chain with different memory costs

In an observed Markov chain, conditional on the next symbol the remaining future has the same law for every past. The KL chain rule therefore equates next-symbol and whole-future loss. A fixed chain

allows exact prediction with one state per distinct transition row. The power law below concerns growing alphabets with smaller budgets.

Let $n \geq 4$ be even and fix $0 < \gamma < 1$. Let $g_i = 1$ on the first $n/2$ symbols and $g_i = -1$ on the rest. Define

$$h_i = -1 + \frac{2i-1}{n}, \qquad P_{ij} = \frac{1 + \gamma g_i h_j}{n}, \qquad P^{\leftarrow}_{ij} = \frac{1 + \gamma h_i g_j}{n}. \tag{4}$$

Both matrices are positive and doubly stochastic because the two vectors have mean zero. The stationary distribution is uniform. By (3) their scalar cycle determinants agree for every choice of scalar weights.

Forward, it suffices to remember the sign g_{X_t} of the current symbol. In reverse, the forecast depends on its position h_{X_t} on a grid. With fewer than n states some grid points must share a forecast. The next theorem finds the best such grouping.

Write

$$\phi(x) = \tfrac{1}{2}\big((1+\gamma x)\log(1+\gamma x) + (1-\gamma x)\log(1-\gamma x)\big), \qquad w(x) = \phi''(x) = \frac{\gamma^2}{1-\gamma^2 x^2}.$$

For a consecutive block I of indices, let $\bar{h}_I$ be the mean of the h_i over I.

Theorem 3.1. *The forward loss is zero for $L \geq 2$. In the reverse direction, put $K = \min(L, n)$. The minimum excess log loss is*

$$R^{\leftarrow}_{n,L} = \frac{1}{n} \min_{\mathcal{I}} \sum_{I \in \mathcal{I}} \left(\sum_{i \in I} \phi(h_i) - |I| \phi(\bar{h}_I) \right), \tag{5}$$

where $\mathcal{I}$ ranges over partitions into K nonempty consecutive blocks. In particular, exact prediction in the reverse direction requires n states. For $1 \leq L \leq n/2$ the reverse loss is bounded above and below by positive constants depending only on γ times L^{-2}. If $L \to \infty$ and $L/n \to 0$, then

$$L^2 R^{\leftarrow}_{n,L} \;\longrightarrow\; \frac{1}{48} \left(\int_{-1}^{1} w(x)^{1/3}\, dx \right)^3. \tag{6}$$

If $k = n - L \geq 1$ and $k = o(n)$, then $R^{\leftarrow}_{n,L} \sim \gamma^2 k / n^3$. With one state the loss in both directions is $n^{-1} \sum_i \phi(h_i)$.

Proof. The forward predictor stores the current sign. For reverse prediction, write $q(x)_j = (1 + \gamma x g_j)/n$, so symbol i has forecast $q(h_i)$. The best forecast in state s is $q(\mathbb{E}[h_{X_t} \mid S_t = s])$, since q is affine. Summing the relative entropy over the two halves of the alphabet gives

$$D_{\mathrm{KL}}(q(x) \| q(y)) = \phi(x) - \phi(y) - \phi'(y)(x - y).$$

This loss is positive for $x \neq y$. By Section 2, we can assign each point to a single state. For two forecasts with parameters $a < b$, the difference in their costs is affine in x with slope $\phi'(b) - \phi'(a) > 0$. Assigning each point to its cheapest forecast, with ties going to the smaller parameter, therefore makes each group consecutive. Strict convexity shows that splitting a group of two or more points decreases the loss. The optimal partition therefore uses $\min(L, n)$ groups, which gives (5). To compute the minimum, minimize over the endpoints of the last block and solve the same problem on the preceding points. Cumulative sums of h_i and $\phi(h_i)$ give each block's cost. With one state the forecast is uniform in both directions. By stationarity the two conditional entropies agree, so the reverse formula gives the common loss.

For a block I of m points, let J_I be the interval formed by the m grid cells of length $2/n$ centred at those points. Write $\Delta_I = 2m/n$. Taylor's theorem and the squared deviations from the block mean give

$$\frac{1}{n}\left(\sum_{i\in I}\phi(h_i) - m\phi(\bar{h}_I)\right) = \frac{w(\xi_I)}{48}\left(\Delta_I^3 - \frac{4\Delta_I}{n^2}\right) \quad (\xi_I \in J_I). \tag{7}$$

For a block of one point both sides vanish. The L widths sum to two. Jensen's inequality gives $\sum_I \Delta_I^3 \geq 8/L^2$. Using $w \geq \gamma^2$, the lower bound is therefore $\gamma^2(L^{-2} - n^{-2})/6$. Blocks of sizes differing by at most one give the upper bound $O_\gamma(L^{-2})$.

For an optimal partition with $L \to \infty$ and $L/n \to 0$, the same estimates give $\max_I \Delta_I = O_\gamma(L^{-2/3})$. Since w is positive and continuous, we can replace $w(\xi_I)\Delta_I^3$ by $(\int_{J_I} w^{1/3})^3$ with relative error $o(1)$ uniformly in the blocks. The linear terms in (7) together contribute $O_\gamma(n^{-2})$. Jensen's inequality on the block integrals gives the lower limit in (6). For the upper limit, start with L intervals having equal integrals of $w^{1/3}$ and round the endpoints to the nearest grid boundaries. Before rounding, their widths are bounded above and below by constant multiples of L^{-1}. Rounding changes each integral by $O_\gamma(n^{-1}) = o(L^{-1})$, and all blocks remain nonempty for large n. These blocks attain the upper limit.

When $k = n - L$ is small relative to n, the inequality $m(m^2-1) \geq 6(m-1)$ in (7) gives a lower bound of $\gamma^2 k/n^3$. To attain it asymptotically, pair the $2k$ grid points nearest zero into k adjacent pairs and give each remaining point its own state. The curvature on the paired cells is $\gamma^2(1 + o(1))$. □

Appendix B formulates the same grouping principle for next-symbol forecasts of a stationary process.

4 From past observations to future laws

The whole past may affect a hidden-source forecast. Its influence passes through the current distribution over hidden states. We compare whole-future loss with squared error in that distribution, then use this comparison to prove the fixed four-symbol example. The analogous bounds for next-symbol forecasts are in Appendix B.

4.1 Whole-future forecasts in a hidden Markov model

For a hidden Markov source, the past influences every future block through the current distribution over hidden states. Write G for the finite set of hidden states and $T^x_{gg'}$ for the probability of emitting x and moving from g to g'. For $w = x_1 \cdots x_\ell$, put $T^w = T^{x_1} \cdots T^{x_\ell}$. Let Y now be the conditional distribution of the hidden state at time 0 given the past. We call this probability vector the mixed state and write μ for its law on Δ_G. The future law with starting distribution y is $P_y = \sum_g y_g P_g$, where P_g is the future law starting from g. Thus P_Y is the forecast from the whole past. With one hidden state the loss defined below is zero, so take $|G| \geq 2$. For a feature S of the past with at most L values, the loss is

$$d(S) = I(\text{past}; \text{future} \mid S) = \mathbb{E}\, D_\infty\big(Y, \mathbb{E}[Y \mid S]\big), \qquad D_\infty(y,c) = D_{\mathrm{KL}}(P_y \| P_c),$$

Write $d_L = \inf_S d(S)$ over features with at most L values. Here $|\cdot|$ is the Euclidean norm. Positivity gives the upper quadratic bound, and (O) ensures that future words distinguish hidden mixtures.

(P) every entry of every T^x is at least $\tau > 0$;

(O) for some ℓ the vectors $T^w\mathbf{1}$, $w \in \mathcal{A}^\ell$, span $\mathbb{R}^G$, so the distribution of that future block determines the starting mixture.

Under (O) the number $\kappa = \min\{\sum_{w\in\mathcal{A}^\ell} |z \cdot T^w\mathbf{1}| : \ |z| = 1,\ \sum_g z_g = 0\}$ is positive. The process "mess3" of [21] satisfies both conditions.

Theorem 4.1. *Assume* (P).

1. *Almost surely every coordinate of Y is at least τ. For $y, c \in \Delta_G$ with each $c_g \geq \tau$ one has $D_\infty(y, c) \leq \tau^{-1}|y - c|^2$. Under* (O)*, the lower bound $D_\infty(y, c) \geq \frac{1}{2}\kappa^2|y - c|^2$ holds for all $y, c \in \Delta_G$.*
2. *$d_L = \inf_{c_1,\ldots,c_L\in\Delta_G} \int \min_k D_\infty(y, c_k)\, d\mu(y)$.*
3. *Partition Δ_G into cubes of side ρ and let the feature S_ρ be the cube containing Y. Then $d(S_\rho) \leq |G|\rho^2/\tau$. Consequently, if the number of cubes meeting the support of μ is at most $C\rho^{-D}$, a distortion ε is reached with at most $C\,(|G|/(\tau\varepsilon))^{D/2}$ features, and with features of entropy $H(S_\rho)$ at $\rho = (\tau\varepsilon/|G|)^{1/2}$.*
4. *Assume* (O)*. If $\mu(B(y, \rho)) \leq A\rho^d$ for every y in the support and $\rho_0 \leq \rho \leq 1$, then $d_L \geq \frac{1}{16}\kappa^2(2AL)^{-2/d}$ for $1/(2A) \leq L \leq \rho_0^{-d}/(2A)$.*

Proof. The mixed state after one more symbol x is $yT^x/(yT^x\mathbf{1})$. Its numerator has coordinates at least τ by (P) and its denominator is at most one. This gives the first statement. The law P_y is the image of the law y of the hidden state under a fixed Markov kernel, so by the data-processing inequality $D_\infty(y, c) \leq D_{\mathrm{KL}}(y\|c) \leq \sum_g (y_g - c_g)^2/c_g$, which is the upper bound. Under (O), $D_\infty(y, c)$ is at least the divergence of the laws of the first ℓ symbols, which assign the probabilities $y \cdot T^w\mathbf{1}$ and $c \cdot T^w\mathbf{1}$ to w. Pinsker's inequality and the definition of κ give $D_\infty(y, c) \geq \frac{1}{2}\kappa^2|y - c|^2$.

For part 2, the law of the future given $S = k$ is P_{c_k} with $c_k = \mathbb{E}[Y \mid S = k]$, because $y \mapsto P_y$ is affine. Hence $d(S) = \mathbb{E}D_\infty(Y, c_S) \geq \mathbb{E}\min_k D_\infty(Y, c_k)$. Conversely, for given centres let S be an index minimizing $D_\infty(Y, c_k)$. For each ℓ the identity (2), applied to the first ℓ symbols, shows that the conditional law given S has divergence at most that of the forecast P_{c_S}. Letting $\ell \to \infty$ gives the identity for the whole future.

For part 3, the mean of Y over a cube lies in the cube, so it is within $\rho\sqrt{|G|}$ of every point of the cube, and its coordinates are at least τ. Apply the upper bound of the first assertion. For part 4, take $\rho = \frac{1}{2}(2AL)^{-1/d}$. A ball of radius ρ about a centre that meets the support lies in a ball of radius 2ρ about a point of the support. Its mass is at most $A(2\rho)^d = 1/(2L)$, so at least half the mass is farther than ρ from all L centres. The quadratic lower bound gives $d_L \geq \kappa^2\rho^2/4 = \kappa^2(2AL)^{-2/d}/16$. □

The upper bound in part 3 is uniform over forecast lengths. Marzen and Crutchfield [21] use the same cubes and obtain an exponent D from a first-order bound. Here the quadratic loss gives $D/2$. The lower bound matches this exponent if (O) holds and the mass bound holds with $d = D$ at all sufficiently small scales. If the cube partition has entropy asymptotic to $d_1 \log(1/\rho)$, the partition chosen in part 3 has entropy asymptotic to $\frac{1}{2}d_1 \log(1/\varepsilon)$.

When (O) fails, the same dimension bounds apply to the projection of Y onto the space V spanned by the vectors $T^w\mathbf{1}$ over all words. The spans for words of length ℓ increase with ℓ, so a single length suffices to span V. For $z = \mathrm{proj}_V(y - c)$, one has $P_y - P_c = \sum_g z_g P_g$, since the two signed measures agree on cylinders. Write $\lambda = \sum_g P_g$ and $p_g = dP_g/d\lambda$. If $c_g \geq \tau$ for every g, then

$$D_{\mathrm{KL}}(P_y\|P_c) \leq \tau^{-1} \int \Big(\sum_g z_g p_g\Big)^2 d\lambda \leq \tau^{-1}|z|^2,$$

since $\sum_g p_g = 1$ and $\int p_g \, d\lambda = 1$. The block map at the spanning length is injective on V, so Pinsker's inequality gives a lower bound by a positive constant times $|z|^2$. The bounds on the number of features therefore apply to the projected distribution, with the forecast in each cell still given by its mean. A constant projection gives zero loss.

Here a single summary of the past determines the forecast for the entire future. Two related uses of finite memory have different objectives. Yang et al. [29, Appendix F] study relative entropy per symbol after grouping states. Feng and Meng [12, Sections 3–4] study generators that retain a finite state throughout generation and prove a squared quantization lower bound for their answers to recursive queries about future events.

4.2 Different rates for a fixed hidden Markov source

The observed example required a growing alphabet. We now give the fixed four-symbol source and apply Theorem 6.1, proved in Appendix A.1.

Proof of Theorem 1.1. Use hidden states $+1, -1$ in that order. Let A_+, A_- be the labelled matrices for two continuation symbols and B_+, B_- those for two reset symbols. Set

$$(A_+, A_-, B_+, B_-) = \frac{1}{16}\left(\begin{pmatrix} 3 & 1 \\ 2 & 2 \end{pmatrix}, \begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix}, \begin{pmatrix} 3 & 3 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 3 & 3 \end{pmatrix}\right).$$

Their sum is doubly stochastic and reversal transposes the four matrices. The parameters in Section 6 are $p_1 = p_2 = r_1 = r_2 = 1/4$, $t_1 = -t_2 = 1/4$ and $h = 1/2$. The two continuation maps carry $[-1/3, 1/3]$ into disjoint subintervals. The contractions are equal, so Theorem 6.1 gives both rates and their deterministic updates. The scalar determinant equality follows from transposition. The lower bounds use only the first future symbol, whereas the upper bounds control the whole future. Thus the rates also hold for every nonempty forecast block, with constants independent of its length. □

The supports of the two posterior distributions have box-counting dimensions $1/2$ and $1/4$. Covering the supports gives loss bounds of orders L^{-4} and L^{-8}, respectively. The extra power in each exact rate comes from the probability of a long run of continuation symbols.

5 The effect of imperfect resets

5.1 Nearby sources with a common exponent

A rank-one reset erases the previous posterior exactly. If the reset retains even a small dependence on that posterior, two limits become relevant. We first perturb the source and let the memory budget grow with the new source held fixed. We then let the perturbation shrink as the budget grows.

Theorem 5.1. *Every entrywise neighborhood of the source in Theorem 1.1 contains an open set of positive two-hidden-state sources with all four symbol matrices invertible and*

$$d_L^{\rightarrow} \asymp L^{-\beta}, \qquad d_L^{\leftarrow} \asymp L^{-\beta}$$

for a common exponent β in each source. Both upper bounds are attained by deterministic memory updates. The sources can be chosen converging to the original one with $\beta \to 5$.

Appendix A.6.2 constructs the nearby separated sources and uses Theorem 1.2 for their rates.

5.2 Changing only the resets

We next keep the continuation matrices fixed. The symbols that previously reset the posterior now retain a small dependence on it. Their images overlap in forward time, so the preceding lower-bound argument needs a replacement.

Proposition 5.2. *In the fixed source of Theorem 1.1, keep the continuation matrices and replace the resets, in mass and contrast coordinates, by*

$$B_\sigma(\eta) = \frac{1}{4}\begin{pmatrix} 1 & 0 \\ \sigma/2 & \eta \end{pmatrix}, \qquad \sigma = \pm 1, \quad 0 < \eta \le 1/32.$$

There is a set of Hausdorff dimension zero such that for every η outside it, including all rational η in this interval, there is a common exponent $\beta_\eta < 5$ with

$$d_L^{\rightarrow} = L^{-\beta_\eta + o(1)}, \qquad d_L^{\leftarrow} = L^{-\beta_\eta + o(1)} \qquad (L \to \infty).$$

A deterministic update attains the upper bound $O(L^{-\beta_\eta})$ in each direction. Along these parameters as $\eta \downarrow 0$,

$$5 - \beta_\eta \sim \frac{6 \cdot 4^{-1/6}}{\log 2} \mathbb{E}\left[\sum_{\sigma=\pm 1} (1 + \sigma X/2)^{-1/2}\right] \eta^{1/3}, \qquad X = \sum_{j \ge 1} \varepsilon_j 4^{-j}, \tag{8}$$

where the ε_j are independent fair signs. The coefficient is approximately 13.8280.

Appendix A.6.3 uses separation of finite word maps, allowing overlapping posterior sets. For every $0 < \eta \le 1/32$, it also bounds the upper logarithmic reverse exponent by 5, including at possible exceptions in Proposition 5.2.

5.3 Letting the reset error and memory budget vary together

At fixed η, Proposition 5.2 concerns the eventual exponent as $L \to \infty$. How large must L be before the perturbation changes the reverse rate L^{-9}? For this question we need bounds uniform in both L and η.

Theorem 5.3. *For the perturbation in Proposition 5.2, take $0 < \eta \le 1/32$ and $L \ge 2$. For each fixed $M > 0$,*

$$d_L^{\rightarrow}(\eta) \asymp_M L^{-5} \qquad (\eta^{1/3} \log L \le M).$$

In particular this holds for $L \le \eta^{-A}$ as $\eta \downarrow 0$ for every fixed $A > 0$. In reverse,

$$d_L^{\leftarrow}(\eta) \asymp_M L^{-9} \qquad (L\sqrt{\eta} \le M).$$

There is also an absolute lower bound $d_L^{\leftarrow}(\eta) \gg L^{-9} + \eta L^{-7}$. Thus $L^9 d_L^{\leftarrow}(\eta)$ tends to infinity whenever $L\sqrt{\eta}$ tends to infinity. Both upper bounds are attained in order by a deterministic memory with at most L states, updated after each symbol.

In the construction of Appendix A.7.1, a reset saves a coarser copy of the current memory. As more symbols arrive, it gives less space to that copy. Once the reset lies far enough in the past, it keeps just a part of the word observed since then. Two successive resets give the reverse lower bound even for a predictor allowed to reread the entire past. Beyond the threshold, parts of two successive continuation runs affect the forecast at the same precision. Counting their distinguishable pairs gives the reverse exponent in the next range.

Theorem 5.4. *For the source of Proposition 5.2, fix $1/2 < A \le 2$ and take $L = \lfloor \eta^{-A} \rfloor$. Then*

$$\lim_{\eta \downarrow 0} \frac{\log d_L^{\leftarrow}(\eta)}{\log \eta} = \frac{19A + 4}{3}.$$

The same limit holds when memory must update one symbol at a time.

For example, at $L = \lfloor 1/\eta \rfloor$ the reverse loss is $L^{-23/3+o(1)}$, while the forward loss still has order L^{-5}. The proof in Appendix A.7.2 gives the exponent for every sequence $\eta \downarrow 0$. A leading constant and the reverse exponent for $A > 2$ remain open.

6 Prediction rates from words and cycles

The preceding results assume a known source. We now ask whether short observable statistics determine its prediction costs. For the reset family, three-symbol words recover the exponents and their directions. Rank conditions extend recovery to the whole law. Counterexamples show what can fail when those conditions are absent.

6.1 The rates in a two-state reset family

We begin with the family containing Theorem 1.1 and calculate the parameters that govern its exponents.

Let the two hidden states be $g = \pm 1$. The observed labels are two continuations A_1, A_2 and two resets B_σ, $\sigma = \pm 1$. Choose $p_1, p_2 > 0$ with $p = p_1 + p_2 < 1$, set $v = (1-p)/2$, and take $0 < r_i < 1$, $t_1 t_2 < 0$, $p_1 t_1 + p_2 t_2 = 0$, and $0 < h < 1$. The entries for transitions from g to g' are

$$T^{A_i}_{gg'} = \frac{p_i}{2}\big(1 + g'(t_i + r_i g)\big), \qquad T^{B_\sigma}_{gg'} = \frac{v}{2}(1 + \sigma h g).$$

Assume $|t_i| + r_i < 1$. Assume also that there is a closed interval $I \subset [-1, 1]$ containing zero such that both maps $f_i(x) = t_i + r_i x$ send I into itself and $f_1(I)$ and $f_2(I)$ have positive distance. The sum of the labelled matrices is doubly stochastic, with (g, g') entry $(1 + gg' \sum_i p_i r_i)/2$. The stationary hidden law is uniform, and reversal transposes the labelled matrices.

Only the continuation probabilities and contraction factors enter the exponent formula.

Theorem 6.1. *For $a = 2, 4$, let $s_a \in (0, 1)$ be the unique solution of*

$$(p_1 r_1^a)^{s_a} + (p_2 r_2^a)^{s_a} = 1, \qquad \beta_a = s_a^{-1} - 1.$$

Then $\beta_4 > \beta_2$ and

$$d_L^{\rightarrow} \asymp L^{-\beta_2}, \qquad L^{-\beta_4}(\log L)^{-\beta_4/2} \ll d_L^{\leftarrow} \ll L^{-\beta_4} \quad (L \ge 2).$$

The matching order $d_L^{\leftarrow} \asymp L^{-\beta_4}$ holds for every $h \in (0, 1)$ if

$$\frac{\log p_1}{\log r_1} \ne \frac{\log p_2}{\log r_2} \qquad \text{or} \qquad \frac{\log r_1}{\log r_2} \in \mathbb{Q}.$$

For any fixed choice of the other parameters, the set of h where this matching order can fail has Hausdorff dimension at most D, where $r_1^D + r_2^D = 1$ and $D < 1$. All upper bounds can be achieved by deterministic updates of the retained state as each symbol of the past arrives.

The proof is in Appendix A.1.

For instance $p_1 = 3/10$, $p_2 = 1/5$, $t_1 = 1/5$, $t_2 = -3/10$, $r_1 = 1/4$, and $r_2 = 1/5$ satisfy the hypotheses with $I = [-1/2, 1/2]$. The logarithms of the contractions are incommensurate and the two ratios involving p_i differ. Thus the powers are about 5.3307 and 9.6448, with matching bounds at every reset strength. The exponent equation has the same form as for the quantization of self-similar probabilities [14, 15]. In reverse, one must also bound collisions between forecasts whose continuation words have different contraction products.

6.2 Recovering the exponents and their directions

Suppose the source belongs to this family, with unknown parameters, symbol roles and direction of observation. The scalar determinants give an unordered pair of exponents. To assign them to the two directions, we use a measurement that records the order of three labels around a cycle. Choose a d-by-d unitary matrix U_j for each symbol j. Taking the trace still removes the choice of starting point. For an observed chain on $\{1, \dots, n\}$ with transition matrix P, let the (i, j) block of $\mathcal{T}_U(P)$ be $P_{ij}U_j$. The matrix-weighted Token L-function for closed walks is

$$\mathcal{Z}_U(P; z) = \det(I - z\mathcal{T}_U(P))^{-1} = \prod_c \det\bigl(I - z^{|c|}p(c)U(c)\bigr)^{-1}. \tag{9}$$

The product in this weighted cycle formula [8, 7] runs over directed closed walks of least period equal to their length, up to cyclic rotation. It includes loops and allows immediate backtracking. The probability weight $p(c)$ is the product of the transitions along c. The matrix $U(c)$ is the ordered product of the matrices at each visit. A cyclic rotation leaves the determinant factor in (9) unchanged.

Expanding the logarithm at zero gives $\sum_{k\geq 1} z^k \operatorname{tr}(\mathcal{T}_U(P)^k)/k$. A walk of least period ℓ has ℓ distinct starting positions. Its mth repetition contributes $z^{m\ell}p(c)^m \operatorname{tr}(U(c)^m)/m$ after division by $k = m\ell$. Summing these repetitions gives the logarithm of the factor for c. The sum of absolute values of the length-k trace terms before division by k is at most $d \operatorname{tr}(P^k) \leq dn$, which gives absolute convergence for $|z| < 1$.

For a hidden chain, replace the matrix inside the inverse determinant by $I - z\sum_x d_x T^x \otimes U_x$. The variables d_x select the labels in a coefficient. Comparing the probabilities of oriented cycles is a classical test of reversibility [17, Section 1.5]. We will extract the difference between the two orientations of a triangle.

For the stability bounds, take a compact parameter class inside the strict inequalities defining the family, with all probabilities and the quantities $|ht_i|$ bounded below.

Theorem 6.2. *The one-symbol scalar determinants recover β_2, β_4 as an unordered pair. Two-symbol scalar traces and one cubic coefficient with 2-by-2 unitary weights assign the two exponents to their directions. All the quantities used can be computed from observable word probabilities through length three.*

The classification is stable and the exponents change by $O(\eta)$ under errors at most η in those probabilities, for sufficiently small η on a compact class. Direct word counts also recover the exponents from words through length three. Words through length two are insufficient even when the direction is known.

Proof. In coordinates given by total mass and hidden-sign contrast, the row matrices are

$$A_i = p_i \begin{pmatrix} 1 & t_i \\ 0 & r_i \end{pmatrix}, \qquad B_\sigma = v \begin{pmatrix} 1 & 0 \\ \sigma h & 0 \end{pmatrix}.$$

The change of coordinates is a scalar multiple of an orthogonal matrix, so reversal still transposes them. The one-label determinant $\det(I - zT^x)$ has degree two for A_i, with eigenvalues $p_i, p_i r_i$, and degree one for B_σ, with eigenvalue v. This determines the label roles, p_i, r_i, v, and the two roots in Theorem 6.1. Every scalar determinant for the sum of weighted label matrices is invariant under transposition.

Choose any active label A_i and order the two reset labels as B_1, B_2. Put

$$D = \operatorname{tr}(A_i B_1) - p_i v, \qquad \Delta = \operatorname{tr}(A_i B_1 B_2) - \operatorname{tr}(A_i B_2 B_1).$$

In forward time $B_1 B_2 = v B_1$ and $B_2 B_1 = v B_2$, while $\operatorname{tr}(A_i B_1) + \operatorname{tr}(A_i B_2) = 2 p_i v$. Thus $\Delta = 2vD$ forward and $\Delta = -2vD$ in reverse. Here $D = \pm p_i v h t_i \neq 0$. For the three labels choose $U_{A_i} = \mathrm{i}X$, $U_{B_1} = \mathrm{i}Y$, $U_{B_2} = \mathrm{i}Z$, where

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

In the logarithm of the inverse determinant of $I - z \sum_x d_x T^x \otimes U_x$, the coefficient of $z^3 d_{A_i} d_{B_1} d_{B_2}$ is 2Δ. Indeed the two orientations have Pauli traces 2 and -2 and three cyclic rotations each. Every matrix-weighted determinant through degree two is invariant under reversal, since traces of two factors are unchanged by exchanging the factors. Thus degree three is needed for cyclic measurements to assign the directions.

To recover these quantities from word probabilities, choose a pair x, y with $\Pr(xy) \neq \Pr(x)\Pr(y)$. For example the normalized deviation for $A_i B_\sigma$ is $\sigma h t_i$ forward, and the pair table transposes under reversal. Form the 2-by-2 matrix H of probabilities $\Pr(uv')$ with rows $u \in \{\varnothing, x\}$ and columns $v' \in \{\varnothing, y\}$. Its determinant is the pair deviation $\Pr(xy) - \Pr(x)\Pr(y)$. Form H_a with entries $\Pr(uav')$. The factorization in Proposition 6.3 gives $H_a H^{-1}$ as the matrix for symbol a in a common basis, so it can be used in the trace and determinant calculations.

Choose the empirical pair gap with the largest absolute value. On the compact class the selected determinant, the continuation eigenvalue gaps, and $|D|$ are bounded away from zero. To distinguish the two kinds of labels when the word probabilities have errors, choose a threshold below half the uniform lower bound on the smaller continuation eigenvalue. A reset has a zero eigenvalue, and the magnitude of its estimate stays below the threshold for small errors. Inversion, simple eigenvalues, and the root equations in Theorem 6.1 then give the required local Lipschitz bounds. The root derivatives are controlled since $\sum_i (p_i r_i^a)^{s_a} \log(p_i r_i^a)$ is bounded away from zero.

For the direct calculation define $J(w) = \Pr(w) / \prod_j \Pr(w_j) - 1$. Choose an ordered pair xy with $C = J(xy) \neq 0$, and form $(J(xxy)/C - 1,\ J(xyy)/C - 1)$. In forward time this gives $(r_i, 0)$ with $x = A_i$. In reverse it gives $(0, r_i)$ with $y = A_i$. This follows from applying f_i twice and from the reset to zero. In particular all active roles and contractions are recovered, and p_i is the corresponding marginal. The same gap conditions give stability. In either direction, $\Delta = 2v\{\Pr(A_i B_1) - \Pr(B_1 A_i)\}$, so the marked coefficient records a difference already present in the pair probabilities. Those probabilities and the marginals depend on the masses and $h t_i$ and are independent of r_i. A small decrease in one r_i shrinks its image interval about $t_i \in I$, so it preserves inclusion and separation. It changes both roots in Theorem 6.1, which proves the lower bound on word length. □

Number of observations. Both calculations use the same short-word probabilities. On the compact class in Theorem 6.2, either estimates the two exponents to accuracy ε and assigns their directions with probability at least 3/4 using $O(\varepsilon^{-2})$ consecutive symbols. This order is optimal if the class contains a local interval along which a contraction varies with the other parameters fixed. Appendix A.2 proves these bounds.

6.3 Recovering the whole law from three-symbol words

The next result recovers the entire stationary law within a specified class of hidden-state models. This also determines every finite-memory loss and therefore both loss curves. The rank condition is the one used in the word calculation above. Write $p(w) = \Pr(w)$, and let $K = (p(xy))_{x,y\in\mathcal{A}}$ be the table of adjacent pairs.

Proposition 6.3. *Suppose the source has a representation with at most n hidden states and K has rank n. The probabilities of three consecutive symbols determine the entire stationary law within this class. They determine both whole-future loss curves $d_L^{\rightarrow}$ and $d_L^{\leftarrow}$, and hence their exponents when they exist. In the class with at most two hidden states, three symbols always suffice, including when K has rank one.*

Proof. Choose n row letters U and n column letters V so $H = (p(uv))_{u\in U,v\in V}$ is invertible. For a letter x let $H_x = (p(uxv))_{u\in U,v\in V}$. Write $p(V)$ as a row and $p(U)$ as a column. The formula

$$p(x_1 \cdots x_k) = p(V)H^{-1}(H_{x_1}H^{-1}) \cdots (H_{x_k}H^{-1})p(U)$$

follows by factoring $H = MN$ and $H_x = MT^xN$, with rows πT^u in M and columns $T^v\mathbf{1}$ in N. The loss curves are determined by this word law and its reversal.

One hidden state gives a product law. To cover the remaining case with two hidden states, assume K has rank one. Then K is the product of its marginals. The pair deviations are the pairings of the rows $\pi T^x - p(x)\pi$ with the columns $T^y\mathbf{1} - p(y)\mathbf{1}$. The rows lie in the line annihilating $\mathbf{1}$, and the columns lie in the line annihilated by π. The pairing between the lines is nonzero. Since every pair deviation vanishes, one of the two families must vanish identically. In either case, multiplying the labelled matrices gives $p(w) = \prod_i p(w_i)$. □

This reconstruction from pair and triple probabilities is the classical observable-operator method [1, 6]. Three-observation identification is also known for ordinary hidden Markov models with linearly independent emissions and an invertible hidden transition matrix [13]. Here the rank condition is stated for a source that emits on transitions, and the reconstruction uses exact probabilities. The explicit exponent formulae in Theorem 6.2 also give stability under errors in those probabilities.

6.4 Recovering the law from cycles

For two hidden states, the scalar cycle determinant determines the law up to the ambiguity already seen in the fixed example. A cubic measurement can resolve that ambiguity.

Proposition 6.4. *For a positive two-hidden-state source, the scalar cycle determinant with separate symbol weights determines the observed law up to reversal. If the law differs from its reverse, one matrix-weighted cubic coefficient selects the observed direction.*

Proof. The determinant gives all single traces and pair traces. For $X_x = T^x - \frac{1}{2}\operatorname{tr}(T^x)I$ it therefore gives the Gram table $G_{xy} = \operatorname{tr}(X_xX_y)$. The alternating form $F(X, Y, Z) = \operatorname{tr}(X[Y, Z])$ on the three-dimensional space of traceless matrices satisfies $F(X, Y, Z)^2 = -2\det(\operatorname{tr}(X_iX_j))_{1\le i,j\le 3}$ for $(X_1, X_2, X_3) = (X, Y, Z)$. This follows by checking a diagonal traceless matrix and the two off-diagonal matrix units as a basis. If the X_x span at most a plane, all these values are zero. Otherwise the Gram table gives the coordinates in a basis of three labelled matrices, and all values of F are fixed up to one common sign. Transposition reverses that sign, and the marked cubic for the basis triple distinguishes the signs. The ordered triple traces now follow from their differences and from $XY + YX = \operatorname{tr}(XY)I$ for traceless matrices.

The traces through degree three determine all word traces by the two-by-two trace identities [26]. With $Q = \sum_x T^x$ and $\lambda = \operatorname{tr} Q - 1$, its stationary projection is $E = (Q - \lambda I)/(1 - \lambda)$. Hence the traces determine $p(w) = \operatorname{tr}(ET^w)$. Transposing the labelled matrices transposes this projection and reverses each word in the resulting law. Thus the at most two choices give exactly the ambiguity claimed. □

With more hidden states, short cycles determine the law under a condition on a known collection of reset symbols. Word probabilities use the initial distribution π and the final vector $\mathbf{1}$ of a hidden path. A trace closes the path and sums over its starting state. The reset symbols in the next theorem let us recover the missing initial and final information. Each reset has rank one because its next hidden-state distribution is fixed by that symbol.

Theorem 6.5. *Let a source with n hidden states have positive labelled matrices, including n designated resets $R_i = u_i v_i^{\mathsf{T}}$. Suppose the u_i form a basis and the v_i form a basis. Then the ordered traces $\operatorname{tr}(T^{x_1} \cdots T^{x_k})$ for $k \le 3$ determine all labelled matrices up to one simultaneous similarity, among such sources with the same reset labels. Thus scalar and matrix-weighted cycle coefficients through degree three determine both loss curves. The adjacent-pair table has rank n, so the curves are also determined by observed words of length three.*

Proof. Take positive factors and write $C_{ij} = v_i^{\mathsf{T}} u_j$. Scale each pair of factors reciprocally so $C_{1j} = 1$ for $j > 1$. The one- and two-step traces give $C_{ii} = \operatorname{tr} R_i$ and $C_{j1} = \operatorname{tr}(R_1 R_j)$ for $j > 1$. For distinct $i, j > 1$ the triangle gives $C_{ij} = \operatorname{tr}(R_1 R_i R_j)/C_{j1}$. Thus C is known. Let U have columns u_i and V rows v_i^{T}, so $C = VU$ is invertible. For every symbol x,

$$(VT^xU)_{ij} = \frac{\operatorname{tr}(R_i T^x R_j)}{C_{ji}}, \qquad U^{-1} T^x U = C^{-1}(VT^xU).$$

All the denominators C_{ji} are positive. The matrix of the hidden chain is $Q = \sum_x T^x$. Since Q is positive and stochastic, $E = \lim_{k\to\infty} Q^k = \mathbf{1}\pi$. The probability of the word w is $\operatorname{tr}(ET^w)$, which can now be calculated in the recovered basis.

For distinct x, y, z the scalar coefficient in the logarithm of the inverse determinant gives the sum of the two cyclic orientations. One system of unitary weights gives all the differences. Choose distinct real numbers s_x and let ξ_x be the unit vector in the direction $(1, s_x, s_x^2)$. With the Pauli matrices from Theorem 6.2, use $U_x = \mathrm{i}(\xi_{x,1} X + \xi_{x,2} Y + \xi_{x,3} Z)$. Then $\operatorname{tr}(U_x U_y U_z) = 2 \det(\xi_x, \xi_y, \xi_z)$, which is nonzero for three distinct symbols by the Vandermonde determinant. The marked coefficient gives the difference of the two orientations times this trace. For repeated letters the scalar coefficient suffices. Finally, the reset-pair submatrix of K has entries $(\pi u_i) C_{ij} (v_j^{\mathsf{T}} \mathbf{1})$. It is invertible by positivity and the two basis conditions. □

The triangle calculation of C is related to recovering a Gram matrix from products of overlaps of rank-one orthogonal projectors [9]. Here the incoming and outgoing vectors may differ, and triangles with two reset matrices recover every other symbol matrix. The following family shows that the hypotheses can hold with different exponents in the two time directions.

Proposition 6.6. *For each $n \ge 2$ and $0 < P < 1$ there is a family of positive sources with n hidden states satisfying Theorem 6.5. For all sufficiently small $r > 0$, its whole-future loss curves have orders $L^{-\beta_2}$ forward and $L^{-\beta_4}$ in reverse, where*

$$\beta_a = -\log_n(Pr^a), \qquad a = 2, 4.$$

The pair probabilities are independent of r, so two-symbol words do not determine the exponents. Closed-cycle coefficients through degree two also fail to assign the exponents to their directions.

There are n continuation symbols and n reset symbols. At age k after a given reset, the n^k forecasts are separated on the scale r^k forward and r^{2k} backward. Their spacing and the probability P^k of reaching that age give the two exponents. Appendix A.4 proves the bounds and constructs memory updates attaining both orders.

6.5 Limits of recovery from short data

When there are more than two hidden states, positivity alone does not give the rank condition in Proposition 6.3. The following families show that short probabilities can then conceal changes in both exponents. The ambiguity can persist even when short cycle measurements are also given.

Theorem 6.7. *For each $d \geq 2$ there is a four-symbol family of strictly positive stationary sources with $d+1$ hidden states, indexed by sufficiently small $r > 0$. Word probabilities through length d and ordered cycle traces through length $d-1$ are independent of r. Yet the whole-future losses have exact orders*

$$d_L^{\rightarrow} \asymp L^{-[1+2\log_2(1/r)]}, \qquad d_L^{\leftarrow} \asymp L^{-d[1+4\log_2(1/r)]}.$$

At $d = 4$, even the combination of all three-symbol words and all matrix-weighted cubic coefficients leaves both exponents undetermined. At $d = 2$, varying the reset strengths along with r gives two sources with three hidden states, identical triple probabilities, and different exponents.

A continuation first changes one contrast of the hidden state. Further continuations move it through d coordinates before a reset registers the change in an observed probability. Its rate of contraction changes the spacing of forecasts in both directions. Appendix A.5 gives the construction and the matching loss bounds.

Conclusion

For invertible two-state sources, the word weights of Theorem 1.2 give a common exponent under separation. Exact resets can produce different exponents. Cubic measurements can recover the direction that scalar cycle data omit.

Under what conditions does the word sum still give the true exponent when posterior sets overlap in both directions? The reset perturbation gives one family where it does for all but a set of parameters of Hausdorff dimension zero. For three or more hidden states, can invertible, separated sources have different exponents under reversal? Their posteriors lie in a higher-dimensional simplex. Different directions in that simplex can contract by different factors, so a determinant alone does not give the diameter. A narrower question remains even for exact resets in Theorem 6.1. Do the reverse power bounds match at every reset strength when the contraction logarithms are incommensurate and $p_i = r_i^{\kappa}$?

Acknowledgments

OpenAI Codex and Anthropic Claude assisted with developing and checking proofs in Sections 3–6 and Appendices A.1–B, with the numerical code and calculations in Appendix A.3, and with drafting and revision throughout the manuscript. The author is responsible for the statements, proofs, citations and computations.

A Proofs and examples for hidden Markov sources

The proofs are grouped by their main-text results. The reset-family rates begin in Appendix A.1; symmetry and perturbations are proved in Appendices A.6–A.7.

A.1 Prediction rates for the reset family

Proof of Theorem 6.1. We first list the possible forecasts and their probabilities after a reset. This turns the prediction loss into a squared approximation error. We then choose how many continuation letters to retain and count the forecasts that any smaller memory must miss.

Write x for the conditional mean of the hidden sign. Forward, A_i has probability p_i and sends x to $f_i(x)$, and B_σ has probability $v(1+\sigma hx)$ and sends x to zero. In reverse, B_σ has probability v and sends x to σh. The reverse A_i has probability $p_i(1+t_ix)$ and sends x to $r_ix/(1+t_ix)$. The chance of any continuation is p in both directions.

All transition entries have a positive lower bound. The future law is a mixture of the two future laws started from a hidden sign. Data processing bounds the KL divergence between two such mixtures by the KL divergence between their initial hidden distributions, which is at most a constant times the squared difference of their coordinates on the convex hull of the posterior range. In forward time the probability of B_+ is affine with nonzero slope in x. In reverse the probability of A_1 has the same property. The one-symbol Pinsker inequality therefore bounds the whole-future KL below by a constant times squared distance. On each set of histories with a common retained state, the conditional mean gives the best forecast and is the least-squares centre. This reduces both problems to approximating the posterior mean by a finite set of points on the line.

For a continuation word $w = i_1 \cdots i_k$, put

$$P_w = \prod_{j=1}^{k} p_{i_j}, \quad R_w = \prod_{j=1}^{k} r_{i_j}, \quad b_w = f_{i_1} \circ \cdots \circ f_{i_k}(0).$$

There is almost surely a last reset in the past. Listed most recent first, the forward word since the reset gives the atom b_w with mass $(1-p)P_w$. Listed in chronological order, the reverse word and last reset sign give

$$\frac{aR_w}{1+ab_w} \quad \text{with mass} \quad vP_w(1+ab_w), \qquad a = \sigma h.$$

The latter identity follows by iterating the reciprocal update. The factors in the reverse word probability telescope. The masses are comparable to P_w since $|b_w| \le 1$.

Choosing a memory from the continuation words. We use a stopping rule on the binary tree of continuation words. For any fixed weights $\omega_1, \omega_2 \in (0,1)$ and $s > 0$ with $\sum_i \omega_i^s = 1$, stop a word at the first prefix u with $\omega_u = \prod_{i \in u} \omega_i \le \delta$, where $0 < \delta < 1$. The stopping words partition the infinite words under the product law with probabilities ω_i^s. Their weights satisfy $\min_i \omega_i \, \delta < \omega_u \le \delta$, so there are $\asymp \delta^{-s}$ stopping words. There are also $O(\delta^{-s})$ nodes with $\omega_u > \delta$, since every internal node has two children.

For the forward rate use $\omega_i = p_i r_i^2$. Keep words that end before stopping exactly, and group the rest by their stopping prefix. A group for u has mass P_u and diameter $O(R_u)$, so the total error is $O(\delta^{1-s_2})$. For the converse, use the atom at each stopping word. Distinct stopping words are prefix-incomparable. If they have common prefix z, their points are separated by at least a fixed multiple of R_z, by separation of the two image intervals. Thus balls of radii cR_u about these atoms are pairwise disjoint for a fixed $c > 0$. A ball missed by every forecast centre costs at least a constant times $\omega_u \asymp \delta$. Choose δ so that at least $2L$ stopping words occur to obtain the matching lower bound.

For the reverse upper bound, group words by their two letter counts. For each fixed count vector the product R_w is fixed. For words in that class sharing a prefix u, with n remaining letters, the mass is $O(P_u p^n)$ by summing the weights over all n-letter suffixes. The diameter of the reverse means is $O(R_u^2 r_{\max}^n)$, since b_w varies by $O(R_u)$ and $R_w \le R_u r_{\max}^n$. Put $W_u = P_u R_u^4$ and $b = p r_{\max}^2 < 1$. The contribution of the cell is $O(W_u b^n)$.

Merge all words of length at least K into a single state at zero, choosing $K = O(\log(1/\delta))$ with $b^K \le \delta$. The tail error is $O(\delta)$. For each count vector of length less than K and each reset sign, split a prefix with $n \ge 1$ remaining letters only when $W_u b^n > \delta$. A complete word is kept exactly. For a fixed n, there are at most $n+1$ remaining count vectors. The tree bound with weights $p_i r_i^4$ gives at most

$$O\left(\delta^{-s_4} \sum_{n \ge 1} (n+1) b^{n s_4}\right)$$

splits in total. There are $O(K^2)$ roots, and each split adds at most one leaf to the total. Thus the number of cells is $O(\delta^{-s_4})$ and their total error is $O(\delta^{1-s_4})$.

The forward state is updated by prepending the new symbol and stopping again. A prefix that has already stopped has even smaller weight after a new symbol is prepended. The reverse state before the tail stores the counts, reset sign, and the chosen prefix. After a new continuation, the test weight $W_u b^n$ for every already known prefix decreases. If the old word was truncated, the new stopping prefix is already available. If it was stored exactly, the next symbol is appended before applying the test. The tail state stays fixed on a continuation, and a reset starts a new record in either direction. This proves the updating claim. Finally $p_i r_i^4 < p_i r_i^2$ implies $s_4 < s_2$.

Reverse lower bounds. Stop with weights $\omega_i = p_i r_i^4$ at $\delta = e^{-m}$, where m is a positive integer, and write $s = s_4$. Use only the atoms with positive reset sign. For a stopping word w, place an interval of radius cR_w^2 about $x_w = hR_w/(1+hb_w)$. Its atom has mass comparable to P_w. If the interval misses all L forecast centres, the squared error from this word is $\gg P_w R_w^4 \asymp e^{-m}$. We choose $c > 0$ small enough that every point in the interval is comparable to R_w. Intervals for words of the same product R_w are disjoint: at their longest common prefix z, the two b-values are separated by at least a constant times $R_z \ge R_w$, and the derivative of $hR_w/(1+hb)$ has magnitude comparable to R_w. All these comparisons can be made uniform for h in a compact subinterval of $(0,1)$.

For a bound valid for every parameter choice, take the Bernoulli law with letter probabilities $q_i = \omega_i^s$ and stop at the same threshold. Write n_i for the counts, $\lambda_i = -\log \omega_i$, and $N = n_1 + n_2$. The stopped sum $n_1 - q_1 N$ has second moment $q_1 q_2 \mathbb{E}N = O(m)$, by the martingale square identity and the bound $N \le m/\min_i \lambda_i + 1$. Also $\lambda_1 n_1 + \lambda_2 n_2 = m + O(1)$. Together these show that at least half the stopping mass lies among words with $n_1 = q_1 m/(q_1\lambda_1 + q_2\lambda_2) + O(\sqrt{m})$. For each such n_1 there are $O(1)$ choices of n_2. Each word has stopping mass comparable to e^{-sm}, so there are $\gg e^{sm}$ words among these $O(\sqrt{m})$ count pairs. Each forecast centre hits at most one interval per pair. Choose m so that e^{sm} is a sufficiently large constant times $L\sqrt{\log L}$. A fixed fraction of intervals is missed, giving error $\gg e^{-(1-s)m} \asymp L^{-\beta_4}(\log L)^{-\beta_4/2}$.

If $\log p_1/\log r_1 \ne \log p_2/\log r_2$, any intervals containing the same point have comparable R_w and comparable $P_w R_w^4$. Their count differences therefore satisfy two independent linear equations with bounded right-hand sides, from the logarithms of these products. Both count differences are bounded. Each point lies in only $O(1)$ intervals, because words with the same counts have the same product. When the contractions are powers of a common number, comparable products range over only $O(1)$ values and the same argument applies. Taking $e^{sm} \asymp L$ with a sufficiently large constant now gives the matching lower bound for every h.

The exceptional reset strengths. The preceding arguments give matching bounds when the counts or the contraction product limit how many intervals can overlap. The last case needs a bound on the

reset parameters that permit many overlaps. It remains to treat the proportional case, where $p_i = r_i^\kappa$ for the same $\kappa > 0$. Put $\rho = e^{-m/(\kappa+4)}$, so every stopping word has $R_w \asymp \rho$. Let $D \in (0,1)$ solve $r_1^D + r_2^D = 1$. Here $r_1 + r_2 < 1$ because the two images of I are separated inside I. In any interval of length $u \geq \rho$, the number of stopping words with b_w in that interval is $O((u/\rho)^D)$. To see this, stop the interval construction when a contraction product first falls below u. Those prefix intervals have sizes and gaps comparable to u, so only boundedly many can meet the given interval. Under the Bernoulli weights r_i^D each such prefix has mass $O(u^D)$, and each of its stopping descendants has mass comparable to ρ^D. When u is within a fixed factor of ρ, the same bound follows directly from separation of the stopping words.

Fix a compact interval $J \subset (0,1)$ for h. Let ν be a finite measure on J that assigns mass $O(u^\alpha)$ to every interval of length u, for some $\alpha > D$. Suppose two stopping intervals collide for some $h \in J$. Write z for the longest common prefix, $a = R_{w'}/R_w$, and $\Delta = |\log a|$. The two products are comparable. The collision implies

$$|1 - a + h(b_{w'} - a b_w)| \ll_J c\rho, \qquad |b_{w'} - b_w| \ll_J \Delta + c\rho.$$

On the other hand $|b_{w'} - b_w| \gg R_z$, and $R_z \gg \rho$. Choosing c sufficiently small gives $\Delta \gg_J R_z$. In particular the products differ. If $n \in \mathbb{Z}^2$ is the count vector of w' minus that of w, then $|n|_1$ is bounded by the sum of the two suffix lengths after z. Since $R_w, R_{w'} \asymp \rho$, this sum is $O(1+\log(R_z/\rho))$. Thus a collision requires $\rho/\Delta \leq Ce^{-\eta|n|_1}$ for some $C, \eta > 0$.

For fixed w and n, the ratio a, hence Δ, is fixed. There are $O((\Delta/\rho)^D)$ candidates for w' by the interval count, since $b_{w'}$ must lie within $O(\Delta)$ of b_w and $\Delta \gg \rho$. For each candidate the first collision inequality restricts h to an interval of length $O(\rho/\Delta)$, of ν-mass $O((\rho/\Delta)^\alpha)$. Indeed, at any point of overlap the coefficient of h has magnitude comparable to $|1-a| \asymp \Delta$, using the small choice of c. The integral with respect to ν of the fraction of stopping words involved in a collision is therefore bounded by

$$C \sum_{\substack{n \in \mathbb{Z}^2,\ \Delta(n) > 0 \\ \rho/\Delta(n) \leq Ce^{-\eta|n|_1}}} \big(\rho/\Delta(n)\big)^{\alpha - D}, \qquad \Delta(n) = \left| \sum_i n_i \log r_i \right|.$$

For a fixed n the contributing ρ form a geometric tail as m increases. Their sum in this display is $O(e^{-\eta(\alpha-D)|n|_1})$, regardless of the size of $\Delta(n)$. Summing in n shows that the integrals over J are summable in m. By Tonelli's theorem the collision fraction tends to zero for ν-almost every h. The remaining intervals are pairwise disjoint, giving the matching bound at $e^{sm} \asymp L$. The set where the fraction fails to tend to zero is Borel. If it had dimension greater than D, Frostman's lemma would give a nonzero measure ν of the specified kind carried by that set [22, Chapter 8]. This proves the dimension bound on J, and a countable compact cover gives it on $(0,1)$. □

A.2 Sample bounds from observed words

On a compact class, overlapping length-three word frequencies from N consecutive observations have variance $O(1/N)$. For separated blocks, condition on the hidden state after the first block. The hidden transition matrix has second eigenvalue $\sum_i p_i r_i < 1$ uniformly, so the covariances form a summable geometric series. Theorem 6.2 then gives an accuracy ε with probability at least $3/4$ using $O(\varepsilon^{-2})$ observations for either recovery rule.

This order is optimal whenever the parameter class contains a nontrivial local interval varying one contraction factor r_i with all other parameters fixed. Along such an interval the exponent varies with nonzero derivative, by its root equation. A parameter change of size $c\varepsilon$ can therefore

separate two exponent values by more than 2ε. The full hidden and observed N-step laws have KL divergence $O(N\varepsilon^2)$ by the chain rule and strict positivity. Data processing gives the same bound for the observed laws. Pinsker's inequality then rules out a successful test of the two values with probability $3/4$ when N is a sufficiently small multiple of ε^{-2}.

A.3 A numerical check for the fixed source

We enumerate the posterior distribution in Theorem 1.1 through age nine and assign the mass of older histories to the representatives at the cutoff. We minimize squared error on this finite distribution by dynamic programming. For memory budgets $8 \to 16$, $16 \to 32$, and $32 \to 64$, the local log-log slopes are -4.87, -4.94, and -4.97 forward, and -7.94, -8.55, and about -8.8 backward. A cutoff at age ten gives the same values at the displayed precision. The calculation uses squared error in the posterior coordinate, which the proof compares with whole-future relative entropy. The accompanying file `check_fixed_hmm.py` reproduces this finite calculation; it is a numerical check, not part of the proof.

A.4 Independent reset vectors

We prove Proposition 6.6. Work in an orthonormal basis whose first vector is the normalized all-ones vector. A probability row in the new coordinates is proportional to $(1, x^{\mathsf{T}})$, where $x \in \mathbb{R}^{n-1}$ records its contrasts. Take two sets t_i, a_j of n small, affinely independent vectors, with $\sum_i t_i = \sum_j a_j = 0$. Choose them so that for each j the numbers $a_j^{\mathsf{T}} t_i$ are distinct. For example, one can rotate one centered simplex relative to the other to avoid the finitely many orthogonality conditions. Put

$$A_i = \frac{P}{n}\begin{pmatrix} 1 & t_i^{\mathsf{T}} \\ 0 & rI \end{pmatrix}, \qquad R_j = \frac{1-P}{n}\begin{pmatrix} 1 \\ a_j \end{pmatrix}\begin{pmatrix} 1 & a_j^{\mathsf{T}} \end{pmatrix}.$$

In the original hidden coordinates all entries are positive for small vectors and small r, by comparison with the projection onto the constant vector. Their sum has first row and column equal to those of that projection in the new basis. The hidden stationary law is uniform, and reversal transposes each matrix. The reset vectors form a basis since the a_j are affinely independent.

A reset j sends the contrast to a_j and has probability $(1-P)(1+x^{\mathsf{T}}a_j)/n$ in either direction. A continuation has probability P/n and sends x to $t_i + rx$ forward. In reverse its probability is $(P/n)(1+x^{\mathsf{T}}t_i)$, and it sends x to $rx/(1+x^{\mathsf{T}}t_i)$. The probability of any continuation is P, regardless of x, in both directions. The reset probabilities distinguish all contrasts because the a_j span $\mathbb{R}^{n-1}$. By Theorem 4.1, predicting from L states has loss comparable to the minimum mean squared error of grouping the contrast vectors into L cells.

After a reset j and k continuations, the forward contrast is $b + r^k a_j$, where b is the sum of the last k vectors t_i weighted by $1, r, \dots, r^{k-1}$, most recent first. In reverse, b is the same sum in chronological order and the contrast is $r^k a_j/(1 + a_j^{\mathsf{T}} b)$. This follows by transposing the product of the continuation matrices in the opposite order. The probabilities of the corresponding past events are $(1-P)(P/n)^k/n$ forward and this same value multiplied by $1 + a_j^{\mathsf{T}} b$ backward. The latter factor is uniformly bounded away from zero and infinity.

Fix a reset j. If two words first differ at position i in the appropriate order, the difference at that position in $a_j^{\mathsf{T}} b$ is at least cr^{i-1} for some $c > 0$ after subtracting the entire remaining tail, provided r is small. The n^k forward points are therefore separated by $\gg r^k$. Taking the reciprocal of the reverse denominator and multiplying by $r^k a_j$ changes distances by a factor comparable to r^k, giving separation $\gg r^{2k}$. Choose k with $2L \le n^k < 2nL$. At least half the points are separated from every

one of L centres by a constant times the corresponding scale. Their masses give lower bounds of orders $(Pr^2)^k$ and $(Pr^4)^k$.

For the forward upper bound, retain the last k continuations, or the exact word and reset label for shorter ages. For the reverse bound, store the reset label, age and whole continuation word through age k. At age s with $k < s < 2k$ store only its first $2k - s$ continuations. Store one state for all ages at least $2k$ and use zero contrast there. The state counts in each direction are $O(n^k)$ by a geometric sum. The worst contrast errors are $O(r^k)$ and $O(r^{2k})$ respectively and are incurred only at ages at least k, of total probability P^k. These rules update from the retained state and the next symbol, including after a reset. Taking n^k a suitable fixed fraction of L proves the upper bounds. Finally $p(xy)$ uses only the first row of the first matrix and the first column of the second, neither of which depends on r. Cubic order for cyclic direction recovery is necessary here because all closed-cycle coefficients of degree at most two agree under reversal as in Theorem 6.2.

A.5 Models that agree on short words

We prove Theorem 6.7 in the same orthonormal coordinates, now of dimension $d+1$. Let $e_1, \dots, e_d$ be the contrast basis and let S be the cyclic permutation $e_j^{\mathsf{T}} S = e_{j+1}^{\mathsf{T}}$ (indices modulo d). For small $t, h, r > 0$, with $r < 1/4$, set

$$A_s = \tfrac{1}{4} \begin{pmatrix} 1 & s t e_1^{\mathsf{T}} \\ 0 & rS \end{pmatrix}, \qquad B_\sigma = \tfrac{1}{4} \begin{pmatrix} 1 & 0 \\ \sigma h e_d & 0 \end{pmatrix}, \qquad s, \sigma \in \{+1, -1\}.$$

For instance, take each constant less than $1/(4(d+1))$. In the original basis each matrix is then close enough to one fourth of the constant projection to have positive entries. Their sum has contrast block $rS/2$ and both off-diagonal blocks zero, so the hidden law is uniform.

A word probability is the mass-to-mass entry of the product. After leaving the mass coordinate a term needs one A to enter contrast 1, another $d-1$ continuations to reach contrast d, and one B to return to mass. Thus $p(w) = 4^{-|w|}$ through length d. For a product containing a B, rotate to start at one. Since each B has rank one, its trace is $4^{-|w|}$ times the product, over runs of continuations, of factors $1 + \sigma h b^{\mathsf{T}} e_d$, where b is the affine sum for that run. Each factor is one if the run has fewer than d letters. For a product of k continuations alone the trace is $4^{-k}(1 + r^k \operatorname{tr} S^k)$. This is 4^{-k} for $1 \le k < d$, proving the assertion about short cycles.

Whole-future loss is again comparable to squared contrast error. For the lower comparison in forward time the probabilities of $A_+^j B_+$, $0 \le j < d$, have contrast coefficients proportional to $S^j e_d$. In reverse, the probabilities of A_+^j for $1 \le j \le d$, multiplied by 4^j, have contrasts $t \sum_{i=0}^{j-1} r^i x^{\mathsf{T}} (S^{\mathsf{T}})^i e_1$. Successive differences give all the coordinates. These probabilities are marginals of a finite future block, so the lower comparison follows from Theorem 4.1. The representation has minimum dimension $d+1$. Indeed the columns $\mathbf{1}, T^{A_+^j B_+}\mathbf{1}$ for $0 \le j < d$ span that dimension. So do the rows $\pi T^{A_+^j}$ for $0 \le j \le d$: after removing the factors 4^{-j}, successive differences isolate the d rotated contrast vectors. The word probabilities for these row and column words therefore have a submatrix of rank $d+1$, a lower bound on the dimension of any representation.

A forward reset gives zero contrast. At age k the contrast is b with

$$b^{\mathsf{T}} = t \sum_{j=1}^{k} s_j r^{j-1} e_1^{\mathsf{T}} S^{j-1},$$

with signs listed most recent first. The mass of each word at age k, summing over reset signs, is $4^{-k}/2$. The first differing sign dominates the rest of its coordinate, so the 2^k contrasts are separated

by $\gg r^k$. Storing only the most recent k signs and keeping shorter ages exactly gives error $O(r^k)$ on a set of probability 2^{-k}. The separation and these upper bounds prove the forward exponent.

In reverse, each reset sign σ has probability $1/4$ and resets the contrast to $\sigma h e_d$. Write the following continuation signs in chronological order. The contrast row at age k is

$$\frac{\sigma h r^k e_d^{\mathsf{T}} (S^{\mathsf{T}})^k}{1 + \sigma h b^{\mathsf{T}} e_d}.$$

The probability of that reset and continuation word at age k is $4^{-k-1}(1 + \sigma h b^{\mathsf{T}} e_d)$ by multiplying the transposed matrices. The denominator sees only the signs in positions $d, 2d, \ldots$. At $k = dm$ a fixed reset sign thus gives 2^m points with individual masses comparable to 2^{-k-m}. The first differing effective sign separates their denominators by $\gg r^k$, and the points themselves by $\gg r^{2k}$. Choose $2L \le 2^m < 4L$. At least half the points are separated from all L centres by a fixed multiple of r^{2k}. Their masses are comparable, so they account for a fixed fraction of that age and reset mass. This gives loss $\gg (r^4/2)^{dm}$.

For the matching upper bound put $K = dm$. Through age K keep the age, reset sign and all effective signs. At ages $K < k < 2K$ keep only those effective signs in the first $2K - k$ positions. Beyond that range a single state represents zero contrast. The number of states is $O(2^m)$, by summing $2^{\lfloor k/d \rfloor}$ up to K and $2^{\lfloor (2K-k)/d \rfloor}$ for the intermediate ages. At an intermediate age the error in the denominator from discarded signs is $O(r^{2K-k})$, hence the error in the contrast is $O(r^{2K})$. For $k \ge 2K$ the contrast itself has this bound. The error is incurred only on the age tail of mass 2^{-K}. Thus the loss is $O((r^4/2)^{dm})$, proving the exponent. The stored prefixes and ages also update after each symbol.

For $d = 2$ the only triples that differ from the product law are $A_s A_{s'} B_\sigma$. Their probabilities are $4^{-3}(1 + \sigma s h t r)$. Keep h and the product tr fixed while varying r in a small interval. This preserves positivity and all triple probabilities, and changes both exponents. For example, take $h = 1/29$ and $(r, t) = (1/64, 1/96)$ or $(1/128, 1/48)$. At $d = 4$ the same two values of r work with $h = 1/29$ and $t = 1/96$ fixed, giving equal short cycle traces as well.

A.6 Invertible updates and perturbations

A.6.1 The common exponent under time reversal

The proof of Theorem 1.2 has three parts. First, word probabilities and derivatives give an upper bound on the loss. Second, disjoint posterior sets give a matching lower bound in that direction. Third, we transfer the logarithmic lower bound to the other direction, even when its posterior sets overlap.

The weight of a word. Work under the hypotheses of Theorem 1.2. The source has at least two symbols and positive invertible labelled matrices T^a. For a word w listed most recent first, write T_w for the product in chronological order, p_w for the stationary probability of the word in that order, and F_w for the normalized update of the posterior. The composition F_w has the most recent update on the outside. Set

$$R_w = \frac{|\det T_w|}{p_w^2}, \qquad W_w = p_w R_w^2 = \frac{|\det T_w|^2}{p_w^3}.$$

Choose a compact interval $J \subset (0, 1)$ whose interior contains the stationary prior and every image $F_a([0, 1])$. For $y \in J$, the word probability from y is a positive linear function of $(y, 1 - y)$ and is uniformly comparable to p_w. The derivative of the normalized update in the probability coordinate

has absolute value $|\det T_w|$ divided by the square of that same function. Thus all the derivatives have absolute values comparable to R_w, uniformly in w and $y \in J$.

The log odds $u = \log(y/(1-y))$ give an equivalent metric on J. Each symbol acts by a contraction in this metric: for a matrix with entries $a, b, c, d > 0$ in row order, the derivative in log odds has absolute value

$$\frac{|ad - bc|e^u}{(ae^u + c)(be^u + d)} < 1.$$

The supremum is strictly below one, since the derivative tends to zero at both ends. Let K be the support of the posterior. It is the compact set obtained by composing the updates along an infinite past. It is contained in J. Every finite word has positive probability, so each finite prefix of such a past contributes to the support. In particular $K = \bigcup_a F_a(K)$.

For a concatenation in either order the probability of the second block, conditional on the first, is comparable to its stationary probability. Multiplicativity of the determinant now gives $W_{uv} \asymp W_u W_v$, with constants independent of both lengths. For $s \geq 0$ put

$$Z_j(s) = \sum_{|w|=j} W_w^s, \qquad P(s) = \lim_{j\to\infty} \frac{1}{j} \log Z_j(s).$$

Indeed $\log Z_{j+k}(s) - \log Z_j(s) - \log Z_k(s)$ is bounded by a constant depending on s. Applying subadditivity and superadditivity after adding or subtracting that constant gives the limit and $\log Z_j(s) = jP(s) + O_s(1)$. The word weights are bounded above and below by exponentials in j, with a strictly negative upper rate by contraction. Thus P is continuous and strictly decreasing. It is positive at zero, and negative at one because $\sum_{|w|=j} p_w = 1$ and R_w decays uniformly. Let s_* be its unique zero. Normalize the weights $W_w^{s_*}$ on each level and take a subsequential limit of the prefix marginals. For every prefix u, its limiting probability is comparable to $W_u^{s_*}$, by $W_{uv} \asymp W_u W_v$ and $Z_j(s_*) \asymp 1$.

A memory chosen from word probabilities. The observed law is non-i.i.d., so K has positive diameter and at least one one-symbol probability varies with the posterior. In either excluded case the conditional law of the next symbol given the entire past would be constant. The same argument holds in reverse. Theorem 4.1 therefore compares loss to squared posterior error in each direction. Stop a past at its first prefix w for which $p_w \operatorname{diam}(F_w(K))^2 \leq \delta$, using the log-odds diameter. The derivative estimate makes this product comparable to W_w. Invertibility gives a positive lower derivative bound for each symbol on J, so $\operatorname{diam} F_a(K) \asymp \operatorname{diam} K > 0$. Together with the word estimates, this bounds the change in stopping cost above and below when a letter is appended. Thus the stopped products are comparable to δ. The limiting prefix measure shows that there are $\asymp \delta^{-s_*}$ stopping words. One forecast per word gives squared error $O(\delta^{1-s_*})$. In particular, $d_L \ll L^{-(1/s_*-1)}$.

The upper bound admits a deterministic update. Prepending the new symbol can only decrease the stopping product of the old word because its probability is at most the old marginal probability and the new map contracts log odds. The new stopping prefix is therefore available after prepending the symbol to the old prefix.

The lower bound with disjoint posterior sets. If the sets $F_a(K)$ are pairwise disjoint, two stopping words with longest common prefix z have sets separated by $\gg R_z$. This follows from the first-symbol gaps and the minimum derivative of F_z. Their diameters are $O(R_z)$, so the stopping sets padded by a sufficiently small constant times their own diameters are disjoint. A padded set missing all the forecast centres has word mass at least p_w and contributes $\gg \delta$ to squared error. Choose

$\delta \asymp L^{-1/s_*}$ with the constant small enough to give at least $2L$ stopping words. Then $d_L \asymp L^{-(1/s_*-1)}$ in this direction.

Why reversal preserves the word weights. Under stationary reversal the symbol matrices become $\Pi^{-1}(T^a)^{\mathsf{T}}\Pi$, where Π is the diagonal matrix of stationary hidden probabilities. The probability and determinant of each word then equal those of the reversed word in the original source. Both R_w and W_w are preserved in this correspondence, so the exponent in the upper bound is common to both directions.

Separation of word maps under reversal. We say word maps are exponentially separated on J if the supremum distance between distinct labelled compositions of length n is at least c^n for a fixed $c > 0$ at all sufficiently large n. This property passes to the reversed source. The same holds when the words are built from a fixed block alphabet and the reversed alphabet consists of its reversed blocks.

To prove this, let $S = \Pi^{1/2}$ and use an orthonormal basis whose first vector is $(\sqrt{\pi_1}, \sqrt{\pi_2})^{\mathsf{T}}$. In this basis the first entry of ST_wS^{-1} is p_w, and the corresponding reversed matrix is its transpose. Divide by p_w and write the resulting matrix as $\left(\begin{smallmatrix} 1 & b \\ c & d \end{smallmatrix}\right)$. Positivity gives uniform bounds on the entries, since p_w is a positive weighted sum of the original entries. The absolute determinant R_w has a lower bound exponential in the word length. The posterior coordinates in this basis are mass one and a contrast x with $x = 0$ at the stationary prior. On either invariant interval J, the denominators of the projective maps are bounded away from zero by the word probability comparison. For the reverse map $F(x) = (c + dx)/(1 + bx)$, three evaluations at $-r, 0, r$ recover its entries:

$$c = F(0), \qquad b = \frac{2c - F(r) - F(-r)}{r(F(r) - F(-r))}, \qquad d = \frac{(1 + br)F(r) - c}{r}.$$

Choose $r > 0$ fixed with these points in the reverse interval. The derivative comparison gives $|F(r) - F(-r)| \gg R_w$. If two reverse maps are within ε in supremum norm, their entries differ by at most $C^n\varepsilon$ for some $C > 1$. In the forward maps the same estimate holds, with a fixed factor from their denominators. This proves the transfer.

A lower bound allowing overlap. For the remaining lower bound, we group a fixed number of symbols into one block and place an auxiliary independent law on these blocks. The original source still supplies the probabilities used in the loss. We use block alphabets $\mathcal{W}_m$ of arbitrarily large fixed length m, closed under the correspondence with reversed blocks. The following criterion will let us compare the two laws. Suppose the block matrices have positive determinants, their maps have no common fixed point in either direction, and their compositions are exponentially separated on J. Suppose also

$$Z'_m = \sum_{w \in \mathcal{W}_m} W_w^{s_*} \asymp 1, \qquad \sum_{w \in \mathcal{W}_m} \sup_{x \in J} |F'_w(x)| = O(1).$$

We show that the logarithmic loss exponent is $1/s_* - 1$. Write $s = s_*$. Set

$$q_w = W_w^s / Z'_m, \quad H = -\sum_w q_w \log q_w, \quad U = -\sum_w q_w \log p_w, \quad V = -\sum_w q_w \log R_w, \qquad w \in \mathcal{W}_m.$$

Place independent probabilities q_w on the blocks and let ν be the law of the resulting limit of the block maps. Its mean logarithmic contraction per block is $\chi = -\sum_w q_w \int \log |F'_w(x)| \, d\nu(x)$. The derivative comparisons give $|\chi - V| = O(1)$ uniformly in m. Also $H = s(U + 2V) + O(1) \asymp m$, since $Z'_m \asymp 1$ and the word costs decay exponentially. If ℓ_w is the Lipschitz constant of a block map,

then $H + \sum_w q_w \log \ell_w \le \log \sum_w \ell_w \le C_0$ for a constant C_0. Since $\chi \ge -\sum_w q_w \log \ell_w$, we have $\min(H, \chi) = H + O(1)$.

Apply Rapaport's dimension theorem [24, Definition 1.1 and Theorem 1.2] to the independent block law q, not to the original source law. In log odds the block maps are real analytic on a compact invariant interval, have $0 < |F'_w| < 1$, and have no common fixed point. Rescale this interval to $[0, 1]$. The earlier reversal argument transfers exponential separation of maps; separation at every sufficiently large length implies Rapaport's infinitely-many-lengths condition. His theorem gives $\dim \nu = \min(1, H/\chi)$. Feng–Hu [4, Theorem 2.8] gives exact dimensionality for this independent, conformal system, so the local dimension equals $D = \min(1, H/\chi)$ at ν-almost every point. Related independent-block approximations appear in [24, proof of Corollary 1.3].

This step uses Rapaport's preprint. We do not rely on the earlier Furstenberg-measure argument of [3]: the authors' 2018 clarification [5] identifies an incomplete entropy-porosity step. The direct bounds for a separated direction above are independent of this dimension input.

We now use this dimension estimate to count forecasts that remain separated. Fix a small $\epsilon > 0$. By the law of large numbers and the ergodic theorem, a positive-probability set of codes has the following properties for all sufficiently large n. The first n blocks have original word probability at least $e^{-n(U+C+\epsilon)}$ and map J to an interval of diameter at most $r_n = e^{-n(\chi-\epsilon)}$. Here C is independent of m: the probability of a concatenation differs from the product of its block probabilities by at most a fixed factor per block. The diameter bound follows by summing the logarithmic derivatives on successive tails and using bounded distortion. By restricting the set of codes further while keeping positive probability, their projected points x satisfy $\nu(B(x, r)) \le r^{D-\epsilon}$ for every sufficiently small r, with a common cutoff.

A maximal set of projected points with spacing $6r_n$ has at least ce^{nH_ϵ} elements, where $H_\epsilon = (\chi - \epsilon)(D - \epsilon)$. This follows by covering the set with the $6r_n$-balls and using the last mass bound. The corresponding word intervals lie within r_n of those points, so their r_n-neighborhoods are disjoint. Choose n so that e^{nH_ϵ} is a sufficiently large constant times L up to a bounded factor. Any L centres miss the r_n-neighborhoods of at least L intervals. Their words are distinct, since their diameters are at most r_n. Each disjoint word event contributes at least its probability times r_n^2 when its interval is missed. Thus the loss is bounded below by a constant times $Le^{-n(U+C+\epsilon+2(\chi-\epsilon))}$. On taking $L \to \infty$ and then $\epsilon \to 0$, the upper limit of the loss exponent is at most $(U + C + 2\chi)/\min(H, \chi) - 1 = 1/s - 1 + O(1/m)$. Let $m \to \infty$. The upper loss bound from the stopping rule gives the opposite inequality for the lower limit. That rule requires no separation of posterior sets for its upper bound.

Separation in one direction. For the last assertion of Theorem 1.2, suppose one direction has disjoint sets $F_a(K)$. Word maps in that direction are exponentially separated. Evaluate two compositions at any point of K. After removing their common prefix, the outputs are in different first-symbol sets. The minimum derivative of the common prefix bounds how much the resulting gap can shrink. Word-map separation thus holds in both directions by the preceding transfer.

For even m take all words of length m with positive determinant. If the letters have the same determinant sign this includes every word. If both signs occur, extend each $(m-1)$-word by a fixed letter of the sign needed to make the product positive. The comparison for W_{uv} then proves $Z'_m \asymp 1$. Because m is even, $(\det T^a)^m > 0$ for every symbol, including when the signs of the determinants differ. Our alphabet therefore contains every constant word a^m. The block maps have no common fixed point in either direction: F_a^m and the contraction F_a have the same unique fixed point. A common fixed point of all F_a would give a singleton posterior support. To bound the sum of derivatives, work first in the separated direction. In log odds, take a sufficiently small open neighborhood of K consisting of the points at distance less than a fixed ε from K. It contains each of

its symbol images by contraction, and these images are pairwise disjoint by the gaps between $F_a(K)$. It follows that all length-m images of the neighborhood are disjoint. Their lengths are comparable to R_w times its length. Thus $\sum_{|w|=m} R_w = O(1)$, also in the other direction since reversal preserves the multiset of R_w. The sum of the Lipschitz constants of the selected maps is bounded by the derivative comparison. Every hypothesis of the block criterion has now been checked, so it gives the logarithmic exponent in both directions. This completes the proof of Theorem 1.2.

Proof of Proposition 2.1. Nearby positive models share a compact interval J and a uniform contraction factor $\theta < 1$ in log odds, even at rank-one symbols. Approximation by finite pasts therefore gives weak continuity of the posterior law. By the KL chain rule the jth term of the future loss is $O(\theta^{2j})$ uniformly on J because the updated posteriors differ by $O(\theta^j)$ and all symbol probabilities are bounded below. Thus the whole-future loss is continuous jointly in the model and its two starting posteriors. Centres may be chosen as conditional means in J. For L centres the integral of the minimum loss is jointly continuous by uniform continuity and weak convergence, and minimizing over J^L proves continuity of d_L.

A.6.2 Separated sources near the fixed example

To prove Theorem 5.1, we construct nearby sources with disjoint posterior sets in both directions and evaluate their limiting exponent. Use the orthogonal mass and contrast coordinates of Theorem 6.2. With $\sigma = \pm 1$ let

$$A_\sigma = \tfrac{1}{4}\begin{pmatrix} 1 & \sigma/4 \\ \sigma a & 1/4 \end{pmatrix}, \qquad B_\sigma = \tfrac{1}{4}\begin{pmatrix} 1 & \sigma b \\ \sigma h & bh + \eta \end{pmatrix}.$$

Take $a, b > 0$ small, h close to $1/2$, and $\eta \geq 0$ small. The four original-basis matrices are strictly positive by continuity, and their sum is doubly stochastic since its mass–contrast form is diagonal with first entry one. The determinants of the A labels are $(1/4 - a/4)/16$, and the determinants of the B labels are $\eta/16$. When $a, b, \eta \to 0$ and $h \to 1/2$ the matrices converge to those of Theorem 1.1.

First take $\eta = 0$. In the forward contrast interval $[-2/3, 2/3]$, the two A maps have separated images of opposite sign and bounded away from zero. Both B maps are constant, with distinct outputs $\pm b$ in between. In reverse the B maps are constant with outputs $\pm h$. The reversed A maps are

$$G_\sigma(x) = \frac{\sigma a + x/4}{1 + \sigma x/4}.$$

They contract on $[-2/3, 2/3]$ with Lipschitz constants below $1/2$, and their images lie between the two B values. On the smaller interval $[-2a, 2a]$ they have disjoint images of opposite sign and map the interval into itself. Let C be the compact set of limits of compositions using only the two G maps. It has Lebesgue measure zero because their Lipschitz constants sum to less than one.

The full reverse posterior support consists of C and the points $G_w(\sigma h)$ with w a finite word of A symbols, including the empty word. This set is compact because the distance of $G_w(\sigma h)$ to C tends to zero uniformly as $|w| \to \infty$. The two G images of C are disjoint. For all but a null set of h in any sufficiently small interval about $1/2$, the two G images of the full support are also disjoint. To prove this, compare finite words with different first symbols. If their arguments are both h, or both $-h$, their rational functions are different: at argument zero they lie in opposite images of $[-2a, 2a]$. If the arguments have opposite signs, the functions have opposite strict monotonicity in h. In either case their coincidence set is finite. A point from a finite word meets the opposite image of C on a null set of h, since the map from h is smooth with nonzero derivative on the interval. There are countably many of these tests. This excludes all possible intersections, while the two B values are separated from each other and from both G images.

For such an h every pair of different first-symbol sets has a positive gap, by compactness. The updates in both directions remain uniform contractions on the displayed larger intervals for sufficiently small changes in the matrices. The compact posterior sets and their first-symbol images vary continuously: if a list of contractions with common constant $\theta < 1$ changes by at most ε uniformly, its coding limits change by at most $\varepsilon/(1-\theta)$. Therefore the separation persists when $\eta > 0$ is sufficiently small, and on an open neighborhood of each resulting positive invertible model. In a general nearby model the stationary law and reversed matrices also vary continuously by strict positivity. One-symbol observability holds in both directions since a forward B probability and a reverse A probability vary with the contrast. This property, too, persists under small changes. Theorem 1.2 now gives a common exponent and the upper bounds with deterministic memory updates.

For an explicit example, take $a = b = 1/100$, $h = 1/2$ and $\eta = 1/1000$. Starting with $[-2/3, 2/3]$, form the unions of images after each of up to four symbols. Comparing intervals with different most recent symbols gives a forward gap of $9/500$ at depth one and a reverse gap of $5652184/90012033755$ at depth four. These comparisons use only the rational endpoints, and the unions contain the posterior sets. Taking $b = 0$ with the other parameters fixed gives an example separated in reverse only. The depth-four reverse gap is then $2296297/36489726552$, and both forward reset maps fix 0, so their posterior images meet there.

To see the limiting exponent, in the forward direction the conditional probability of A_σ is $(1 + \sigma a x)/4$. The single-letter factors $|\det A_\sigma|^2 / \Pr(A_\sigma \mid x)^3$ are bounded between u_- and u_+, both tending to $1/64$. For either B the factor is at most $c\eta^2$, since the symbol probabilities stay bounded below. The chain rule factors W_w into these single-letter quantities along the successive posterior updates. Hence for each fixed $s > 0$,

$$\log(2u_-^s) \le P(s) \le \log\bigl(2u_+^s + 2(c\eta^2)^s\bigr).$$

The lower bound counts A-only words. Both bounds tend to $(1 - 6s)\log 2$, so $s_* \to 1/6$ and the common loss exponent tends to 5. This proves Theorem 5.1.

A.6.3 Changing only the reset matrices

We keep the continuation matrices fixed throughout this subsection. First we bound the reverse exponent for every reset parameter. We then prove the common-exponent statement of Proposition 5.2 by checking the block criterion from Appendix A.6.1. Finally we evaluate the change in the exponent as the reset error tends to zero.

A reverse lower bound for every parameter. Take $a = b = 0$, $h = 1/2$ and $0 < \eta \le 1/32$ in the same matrices. Put $r = 1/4$. A reverse block consisting of B_+ followed by k continuations acts by

$$x \longmapsto G_w(x) = \frac{r^k(h + \eta x)}{1 + (h + \eta x) b_w}, \qquad |b_w| \le 1/3,$$

where b_w is the signed sum from the proof of Theorem 6.1. Its conditional probability is $4^{-(k+1)}(1 + (h + \eta x) b_w) \ge (3/4) 4^{-(k+1)}$. Each block sends $[-1, 1]$ into $J_k = [-r^k, r^k]$. The 2^k centres $G_w(0)$ have gaps at least $(12/49) r^{2k}$. Indeed the first differing sign in the sum b_w dominates the tail and gives a gap of at least $(4/3) r^k$, and $b \mapsto r^k h/(1 + hb)$ has derivative of magnitude at least $(9/49) r^k$. The displacement on J_k is at most $(3/2)\eta r^{2k}$ and the derivative there is at least $\eta r^k/2$. Thus the block images are separated on J_k.

Condition on a fixed block to enter J_k, then take n blocks from the 2^k choices. As in the stopping-set lower bound, the 2^{kn} sets have gaps at least a constant depending on k, η times $(\eta r^k/2)^n$. Each has word mass at least a fixed constant times $((3/4)4^{-(k+1)})^n$. Choose n with $2L \le 2^{kn} < 2^{k+1}L$ and count the sets missed by L centres. This gives

$$d_L^{\leftarrow} \gg_{k,\eta} L^{-5-k^{-1}\log_2(64/(3\eta^2))}.$$

Since k is arbitrary, the upper limit of $-\log d_L^{\leftarrow}/\log L$ is at most 5 for every such η. This argument controls the reverse exponent with the continuation matrices held fixed.

Distinguishing the word maps. In contrast coordinates, the forward maps are

$$A_\sigma(x) = (x+\sigma)/4, \qquad B_\sigma(x) = \frac{\eta x}{1+\sigma x/2}.$$

In reverse these become $x/(4+\sigma x)$ and $\sigma/2 + \eta x$, respectively. In both directions the four maps preserve $J = [-2/3, 2/3]$ in the contrast coordinate (equivalently, $[1/6, 5/6]$ in probability coordinates). They have nonzero derivatives and no common fixed point. The forward reset probabilities and reverse continuation probabilities vary with x. The sums of their Lipschitz constants on J are at most $1/2 + 9\eta/2$ forward and $18/25 + 2\eta$ in reverse, both less than one.

For rational η, we first distinguish all finite word maps. Choose a prime ℓ dividing the reduced denominator of η and write v_ℓ for its valuation on the rationals. On $J \cap \{v_\ell(x) < 0\}$ every forward map keeps the valuation negative. For an A this follows because adding ± 1 preserves the valuation and dividing by 4 cannot increase it. For a B the resulting valuation is $v_\ell(\eta)$ if ℓ is odd, and $v_2(\eta) + 1 < 0$ if $\ell = 2$ is the only prime in the denominator. Use an odd prime whenever one is available. On the real interval J, the two A images are disjoint and both avoid the B images. The two B signs can be distinguished by

$$\frac{B_\sigma(x)}{2\eta} = \frac{1}{\sigma + 2/x} \equiv \sigma \begin{cases} \pmod{\ell}, & \ell \text{ odd}, \\ \pmod 4, & \ell = 2. \end{cases}$$

In the second case $v_2(2/x) \ge 2$. Start at the rational point $1/\ell$ in J. The output of any nonempty word determines its outermost symbol. Inverting that map and repeating distinguishes any two different words of the same length. Transposing the word matrices, which also reverses their order, gives the same distinctness in reverse.

For each fixed rational parameter, clear denominators in the four generators using integer scalar multiples. Word matrices of length n have entries of size at most C^n for a fixed $C > 1$. Two distinct fractional linear maps differ at one of any three fixed distinct rational points of J, since the numerator of their difference is a nonzero polynomial of degree at most two. The rational values at such a point differ by at least c^n for some $c > 0$. This gives exponential separation of the maps in supremum norm, at every length. It also gives separation of the maps obtained by grouping a fixed number of letters into one symbol. The grouped families still have no common fixed point, by using the repetitions of individual letters.

For real parameters, consider blocks of a fixed length $m \ge 2$ that contain both types of letter. We show that no two infinite codes built from these blocks give the same analytic function of η. The forward maps are holomorphic and uniformly contracting on $|x| \le 3/4$ for $|\eta| < 1/10$. Their limits along an infinite code are therefore analytic on $|\eta| < 1/10$.

An initial run of k continuations with signs $\sigma_1, \dots, \sigma_k$, listed from outside to inside, sends the following value x_B to $x_A = (N + x_B)/4^k$, where $N = \sum_{i=1}^k \sigma_i 4^{k-i}$ is odd. A run of j resets applied

to a continuation value x_A gives

$$\frac{2\eta^j}{x_B} = \frac{2}{x_A} + \sum_{i=1}^{j} \sigma_i \eta^{j-i}. \tag{10}$$

These formulas show that x_B has coefficients in $2\eta\mathbb{Z}_2[[\eta]]$ and $1/x_A$ has coefficients in $4\mathbb{Z}_2[[\eta]]$, where $\mathbb{Z}_2$ is the ring of 2-adic integers. Indeed, start with finite codes cut after a continuation run and take the innermost input to be 0. Proceeding outwards, $N + x_B$ has odd constant term, as does the right side of (10). This proves both coefficient bounds for the finite codes. Each outer reset increases the order of agreement of two input series by one, since $B_\sigma(x) - B_\sigma(y)$ is $\eta(x - y)$ divided by $(1 + \sigma x/2)(1 + \sigma y/2)$. Every fixed Taylor coefficient therefore stabilizes as the cutoff increases, since the code has infinitely many resets.

The constant term is nonzero precisely for codes starting with a continuation. For such a code, the reduced denominator 4^k and numerator N identify the entire first run: the signs are unique by the largest differing power of 4. A code starting with resets has order of vanishing j by (10). The coefficients of $2\eta^j/x_B$ modulo 8 recover all j signs because $2/x_A$ is divisible by 8 coefficientwise. Now remove the identified run by injectivity at positive η. Repeating shows that the analytic limit identifies the whole code.

Apply [28, Lemma 2.7 and Theorem 2.10] on $V = (-3/4, 3/4)$ and a compact parameter interval in $(0, 1/32]$. The denominators are at least $5/8$ on V. The maps preserve V and their derivatives are uniformly bounded in absolute value by $\max(1/4, 2/25) < 1$. The coding argument proves the nondegeneracy of Definition 2.9 there. Outside a set of Hausdorff dimension zero, the theorem gives exponential separation at $x = 0$ for all sufficiently long block words with different first blocks. The finitely many shorter differing suffix pairs have positive supremum distances, since an exact coincidence would contradict the long-word bound after extension. A common prefix reduces these distances by at most an exponential factor, by its minimum derivative. This gives the required all-pairs separation in supremum norm.

The transfer above gives the same separation in reverse, since reversing a word preserves the block alphabet. In both directions some block maps have disjoint images, by choosing the outermost signs of continuations in forward time and of resets in reverse time. In particular there is no common fixed point. A countable union over lengths and compact parameter intervals gives the exceptional set in the proposition. The restriction to blocks containing both types of letter follows the use of restricted alphabets in Solomyak [27, Section 2].

We now use this separation of maps to bound the loss from below. Let $s = s_\eta$ be the common zero of P. For a fixed block length $m \geq 2$, let $\mathcal{W}_m$ consist of all blocks in the rational case and of the blocks containing both types of letter otherwise. In either case $Z'_m = \sum_{w \in \mathcal{W}_m} W_w^s \asymp 1$ as $m \to \infty$. For the restricted case the upper bound follows from $Z_m(s) \asymp 1$. To get the lower bound, append a fixed two-letter block containing one letter of each type to every word of length $m - 2$ and use $W_{uv} \asymp W_u W_v$. These blocks have total Lipschitz constant at most one, since the sum of the individual Lipschitz constants is less than one in each direction. They have exponentially separated word maps in both directions. The block criterion from the proof of Theorem 1.2 now applies in each direction.

The change in the common exponent. For a continuous function f on J, let $Mf(x) = (f(A_+(x)) + f(A_-(x)))/2$, and let ν_0 be the stationary law of these two affine maps used with equal probabilities. For s close to $1/6$, set $\lambda_0 = 2 \cdot 64^{-s}$, $\delta = (\eta^2/4)^s$, and

$$g_s(x) = \sum_{\sigma=\pm 1} (1 + \sigma x/2)^{-3s}, \qquad Qf(x) = \sum_{\sigma=\pm 1} (1 + \sigma x/2)^{-3s} f(B_\sigma(x)).$$

The operator $\mathcal{T} = \lambda_0 M + \delta Q$ satisfies $Z_n(s) = (\mathcal{T}^n 1)(0)$ in the forward direction. This follows by the chain rule for word probabilities: the single-letter factors of W are $1/64$ for an A and $(\eta^2/4)(1+\sigma x/2)^{-3}$ for a B. At $s = 1/6$ the contribution from the two continuations is one, and $\mathcal{T}1$ exceeds one uniformly on J. By iteration $P(1/6) > 0$, so $s_\eta > 1/6$ for every η in the interval.

For a Lipschitz function, $M^k f$ tends uniformly to $\nu_0(f)$ with error $O(4^{-k})$ times its Lipschitz constant, by coupling two initial points under the same affine maps. Therefore

$$u = \lambda_0^{-1} \sum_{k \geq 0} M^k (g_s - \nu_0(g_s))$$

is uniformly bounded for s near $1/6$. With $f = 1 + \delta u$ we have $\mathcal{T} f = (\lambda_0 + \delta \nu_0(g_s)) f + O(\delta^2)$ uniformly. Indeed the terms linear in δ cancel by the defining series for u. For small η, f is bounded above and away from zero. Positivity of $\mathcal{T}$ lets us iterate the resulting upper and lower inequalities for $\mathcal{T} f / f$. Thus

$$e^{P(s)} = \lambda_0 + \delta \nu_0(g_s) + O(\delta^2).$$

It follows that $s_\eta \to 1/6$. At the root the same estimate gives $s_\eta - 1/6 = O(\eta^{1/3})$, so $(\eta^2/4)^{s_\eta} \sim 4^{-1/6} \eta^{1/3}$. Expanding λ_0 and $1/s_\eta - 1$ at $s = 1/6$ proves the claimed coefficient formula. The measure ν_0 is the law of $\sum_{j \geq 1} \varepsilon_j 4^{-j}$ with independent fair signs.

A.7 Prediction as the reset error decreases and the memory grows

A.7.1 When the original rates persist

The estimates in this subsection are uniform in the reset error and the memory budget. For Theorem 5.3, one reset and the following run give the first lower bounds. A memory that stores a shortened account of earlier runs gives the upper bounds. Two successive runs give the additional reverse lower bound.

In either time direction, let D_L be the least expected squared error in approximating the posterior hidden-sign mean x by L centres. Uniformly in η, $d_L \asymp D_L$. Indeed, the posteriors lie in $[-2/3, 2/3]$, so data processing bounds the future log loss above by a constant times squared error. The probability of B_+ forward, or A_+ backward, is affine in x with a fixed nonzero slope, giving the lower bound by Pinsker's inequality. Conditional means give the optimal forecasts on cells.

Group histories by their last reset B_σ and the following k continuations w. Such a reset exists almost surely since the conditional probability of a continuation is $1/2$. Put $r = 1/4$ and $h = 1/2$. If x is the posterior just before this block, the final posterior $H_{\sigma,w}(x)$ is, respectively,

$$b_w + r^k \frac{\eta x}{1 + \sigma h x}, \qquad r^k \frac{\sigma h + \eta x}{1 + (\sigma h + \eta x) b_w}.$$

In each formula the b_w run over the signed base-four sums $\sum_{j=1}^{k} \epsilon_j 4^{-j}$ (the word order is reversed between the formulas). They lie in $[-1/3, 1/3]$ and distinct values are at least a constant times r^k apart. Conditional on x, the block has probability $4^{-(k+1)}$ times $1 + \sigma h x$ forward, or $1 + (\sigma h + \eta x) b_w$ backward, by multiplying its matrices. Both factors are bounded above and below. The stationary mean of x before the block is zero. Hence its marginal probability is exactly the same as at $\eta = 0$ in either direction. This is used only for a suffix beginning with a reset. In both formulas $|H'_{\sigma,w}| \ll \eta r^k$, and $H_{\sigma,w}(0)$ is the unperturbed value.

For a lower bound, fix the initial x, the sign $+$ and $2L \leq 2^k < 4L$. As w varies the outputs are separated by $\gg r^k$ forward and $\gg r^{2k}$ backward. The second estimate follows by differentiating in b_w, since $h + \eta x$ is bounded away from zero. At least half the outputs miss the L centres by a

fixed fraction of this spacing. Multiplying the squared spacing by their total probability $\gg 2^{-k}$ and integrating over stationary x gives $D_L \gg L^{-5}$ and $D_L \gg L^{-9}$.

A memory that updates after each symbol. Write $b = 5$ forward and $b = 9$ backward. First consider an exact reset. At level n the forward memory keeps the last $\min(k, n)$ letters since the reset and evaluates their base-four sum. The error is at most Cr^n on the event $k \geq n$, of probability 2^{-n}, and is zero otherwise. The reverse memory keeps the sign, the age $k < 2n$ and the *first* $j = \min(k, 2n - k)$ letters since the reset. At $k \geq 2n$ it uses one state with decoded value zero. In the formula for $H_{\sigma,w}(0)$, the first j chronological letters determine b_w to within Cr^j. Setting the missing digits to zero gives zero error for $k \leq n$ and error at most $Cr^{k+j} = Cr^{2n}$ for $n < k < 2n$, whose total probability is at most 2^{-n}. The squared error over $k \geq 2n$ is $O(2^{-10n})$. Thus each direction has squared error $O(2^{-bn})$ using $O(2^n)$ states, since the reverse count sums $2^{1+\min(k,2n-k)}$ over $k < 2n$. Both memories update from one symbol. In reverse, append letters up to age n, then shorten the stored prefix as k increases. In either direction a higher level determines every lower level by discarding letters. Level zero uses a single state.

For each large integer L take level $n = \lfloor \log_2(L/C_*) \rfloor$, where C_* is fixed large enough that this memory, denoted by U_L, uses at most $L/3$ states. Set $K_L = \lfloor \log_2(cL) \rfloor$ for a small fixed $c > 0$. Define the new memory S_L by increasing budget, using one state with estimate zero below a fixed cutoff. Once $k \geq K_L$, retain just U_L. For $k < K_L$, store (σ, w) and the state $S_{m_k(L)}$ held just before the last reset. The allocations below satisfy $1 \leq m_k(L) < L$, decrease weakly with age, and increase weakly with budget. We choose them with $\sum_{k<K_L} 2^{k+1} m_k(L) \leq 2L/3$, so the two kinds of states together use at most L values.

We must also check that a reset can obtain the required earlier state without rereading the past. Inductively, S_L determines S_l for every $l \leq L$. If $k \geq K_L$, coarsen U_L to U_l. If $K_l \leq k < K_L$, compute U_l from the stored sign and word. If $k < K_l$, keep that sign and word and coarsen the stored earlier state to budget $m_k(l)$ by induction. The one-state initial budgets require no data. A reset therefore stores $S_{m_0(L)}$ obtained from the current state. A continuation extends the word and coarsens its earlier state to budget $m_{k+1}(L)$. Upon reaching K_L it initializes U_L from the sign and word, and later continuations update U_L alone. These rules read only the current state and the new symbol. They define a function of a finite past because U_L needs at most $2n + 1$ recent symbols and each reference to an earlier state uses a strictly smaller budget.

Let E_L be the mean squared error of this memory. For $k < K_L$ decode as $H_{\sigma,w}$ applied to the earlier state's estimate. For $k \geq K_L$ use the exact-reset decoder of U_L. The latter contribution is $O(L^{-b} + \eta^2 L^{-5})$ by the common suffix probabilities, $|H'_{\sigma,w}| \ll \eta r^k$ and $\sum_{k \geq K_L} 2^{-k} r^{2k} \ll L^{-5}$. For each shorter suffix the block probability given the entire earlier past is $O(4^{-k})$. Integrating the squared earlier error against its stationary law and summing over the signs and words gives

$$E_L \leq C(L^{-b} + \eta^2 L^{-5}) + C\eta^2 \sum_{k<K_L} 2^{-5k} E_{m_k(L)}.$$

Using the future law started from the decoded mean gives excess log loss at most a constant times E_L, by the mixture bound above. In reverse, induct on L with $E_m \leq Bm^{-9}$ for $m < L$ and take $m_k = \lceil (L/2^k)^{3/5} \rceil$. The choices in both directions give $m_k < L$ for large L, and bounded budgets are covered by increasing B. The number of states allocated to the shorter words is $O((c + c^{2/5})L)$. The weighted sum of their errors is at most $BL^{-27/5} \sum_{k<K_L} 2^{2k/5} \ll BL^{-5}$. Choose ηL^2 smaller than a fixed constant so that the term containing B is at most $BL^{-9}/2$. This proves the claim there. For $L\sqrt{\eta} \leq M$, run the encoder at a smaller budget equal to a fixed fraction of L depending on M. Forward, allow a small change in the exponent to obtain a bound uniform in η. Put $u = 5 - D\eta^{1/3}$ with a large absolute constant D, and first take η small enough that $4 \leq u < 5$. Set $\gamma = 6/(1 + u)$,

$t = 1 - 2^{1-\gamma}$, and $m_k = \lceil Lt\, 2^{-\gamma k}/8 \rceil$. Summing the geometric series shows that these allocations use at most $L/4 + 2cL$ states. Since $5 - u\gamma = \gamma - 1$, the induction hypothesis $E_m \le Bm^{-u}$ bounds the sum in the recurrence by $8^u BL^{-u} t^{-(1+u)}$. Here $t \asymp 5 - u$ uniformly for $4 \le u < 5$. Thus $\eta^2 t^{-(1+u)} \ll D^{-(1+u)} \eta^{(5-u)/3}$, and choosing D large absorbs the term containing B. We obtain $E_L \ll L^{-5+D\eta^{1/3}}$. When $\eta^{1/3} \log L \le M$ this is $O_M(L^{-5})$. For the remaining η in this range, L is bounded in terms of M and the one-state predictor suffices.

The additional cost of two reverse blocks. Take two reverse blocks starting with B_+, of continuation lengths k_1, k_2. Choose $n = k_1 + k_2$ with $2L \le 2^n < 4L$ and choose $d = k_2 - k_1$, of the same parity as n, so that $c_1/16 < \eta 4^d \le c_1$ for a small fixed $c_1 > 0$. Nonnegative lengths exist for $L \ge C_0/\sqrt{\eta}$ with C_0 absolute. After the first block the posterior lies within $O(r^{k_1})$ of zero. For each second word v, the final points therefore lie within $O(\eta r^{k_1+k_2})$ of $H_{+,v}(0)$. These centres have gaps $\gg r^{2k_2}$, and our choice of d makes the radii less than one quarter of the gaps. Within each such group the second map has derivative $\gg \eta r^{k_2}$, while the first block's outputs have gaps $\gg r^{2k_1}$ for fixed initial x. Thus all 2^n final points have gaps $\gg \eta r^{k_2+2k_1}$. Each word pair has conditional probability $\gg 4^{-n}$. Counting the missed points as before gives

$$D_L \gg 2^{-n} \eta^2 r^{2(k_2+2k_1)} = \eta^2 2^{-7n} 4^d \asymp \eta L^{-7}.$$

For smaller L, the earlier L^{-9} lower bound absorbs this term. This proves Theorem 5.3.

A.7.2 The reverse rate beyond the threshold

We next prove Theorem 5.4. The upper bound allocates memory among recent runs, keeping only the digits that can affect the forecast at the required precision. For the lower bound, we count separated forecasts after two runs. Write $r = 1/4$, $h = 1/2$, $H = \log_4(1/\eta)$ and $N = \log_2 L$. In reverse, a reset of sign σ followed by k continuations with signed sum b_w sends the earlier mean x to

$$r^k \frac{\sigma h + \eta x}{1 + (\sigma h + \eta x) b_w}.$$

On the posterior interval the derivatives with respect to x and b_w are $O(\eta r^k)$ and $O(r^k)$, respectively. The same bounds apply if only an initial segment of the digits of b_w is used. As in Theorem 5.3, squared error in the mean and whole-future prediction loss have the same order.

How many digits to keep. Take an integer $H \le n \le 4H$ and put $F = (19n + 8H)/3$. We construct a memory with at most 2^n times a polynomial in n states and squared error at most 2^{-F} times a polynomial in n. Use the seven most recent resets, numbered from newest to oldest. Record their signs and the number k_i of continuations following each, with every count stopped at $K = \lceil F \rceil + 1$. Define

$$S_i = \sum_{l=1}^{i} k_l, \qquad W = \sum_{l=1}^{7} \min(k_l, \lfloor 4H \rfloor), \qquad v = (F - W)/4.$$

For word i keep its first q_i chronological digits, where

$$q_i = \min\{k_i, \max(0, \lceil v - S_i - (i-1)H \rceil)\}.$$

Starting another continuation can increase the length needed for the newest word only if it was kept in full, in which case the new digit is available. The other lengths can only decrease. At a reset, discard the oldest word. This can increase v by at most H, and each surviving word moves one

place farther from the present, adding H to its subtraction in the formula for q_i. Thus every update can be made from the current state and the new symbol. The seven resets exist almost surely in a stationary past. For a finite input stream the same rule applies after its seventh reset.

To count states, put $s_i = \sum_{l \le i} q_l$ and $s_0 = 0$. If $s_7 \ge n$, then $W \ge n$. Indeed the ages sum to at least s_7, and shortening even one contributes $\lfloor 4H \rfloor \ge n$. Hence $v \le v_0 = (4n + 2H)/3$. Whenever $q_i > 0$, using $S_i \ge s_i$ gives

$$2s_i \le s_{i-1} + v_0 - (i-1)H + 1.$$

Otherwise $s_i = s_{i-1}$. Now $s_1 \le v_0/2 + 1 \le n + 1$ and $s_2 \le n + 2$ since $3v_0/4 - H/2 = n$. For $i \ge 3$ we have $v_0 - 2H \le n$. It follows that $s_7 \le n + O(1)$ in every case. There are at most $2^7(K+1)^7$ choices of signs and ages.

If any age reaches K, use zero as the posterior mean. This costs $O(2^{-K})$ in expectation, since each symbol independently has reset type with probability $1/2$. Otherwise apply the seven block maps in time order, starting from zero and using the stored digits. The initial error contributes at most $C\eta^7 r^{S_7} \le Cr^v$, since $F \le 28H$. A word with $q_i = k_i$ adds no error. Any other word contributes at most $C\eta^{i-1} r^{S_i + q_i} \le Cr^v$ by the derivative bounds. Each unsaturated age pattern has probability at most 2^{-W}, so its contribution to squared error is $O(2^{-W} r^{2v}) = O(2^{-F})$. Summation gives the claimed polynomial factor. Its logarithm is only of logarithmic size in the number of retained digits, so it disappears from the exponent limit. For fixed $1/2 < A \le 2$, take $n = \lfloor N - C_0 \log_2(N+2) \rfloor$ with C_0 large. For small η this lies in $[H, 4H]$ and the memory fits within L states. Thus its loss is at most $2^{-(19N+8H)/3 + O(\log N)}$.

Which pairs of words remain distinguishable. Let C_j be the set of j-digit signed base-four sums and C_∞ the corresponding infinite Cantor set. We need one fact about the finite sets. If $f(x) = -\kappa/[h^2(x+\lambda)]$, with $1 \le \kappa \le 4$ and $3/2 \le \lambda \le 5/2$, then for every $\epsilon > 0$ the set $C_j + f(C_j)$ contains at least $c_\epsilon 4^{(1-\epsilon)j}$ points at pairwise distances $\ge c_\epsilon r^j$, uniformly in the two parameters.

To see this, consider first the linear sums $C_\infty + sC_\infty$. They are generated by the four maps $z \mapsto (z + a + sb)/4$, $a, b \in \{-1, 1\}$. No two different infinite addresses give values identical as functions of s, since signed base-four digits have unique addresses in each coordinate. Hochman's parameter theorem [2, Theorems 1.7–1.8] gives Hausdorff dimension one for the sum outside a set of slopes of dimension zero on each compact interval. On the hull of C_∞, $7/6 \le x + \lambda \le 17/6$, so $|f''|$ is bounded above and away from zero uniformly in the parameter rectangle. Thus f' is bi-Lipschitz there and $f'(C_\infty)$ has dimension $1/2$. Choose $x_0 \in C_\infty$ with a nonexceptional slope $s = f'(x_0)$. We now pass from that linear sum to the nonlinear finite sums.

For any sufficiently large fixed D, choose a length m and at least $4^{(1-\epsilon/2)m}$ digit pairs in C_m^2 whose linear sums are $2Dr^m$ apart. Indeed dimension one supplies that many separated points at scale $4Dr^m$ for large m, and truncation moves each by $O(r^m)$. Fix an initial digit string for x near x_0 on whose entire interval $|f' - s| \le r^m/2$, and fix an equally long initial string for y. After these strings concatenate any of the chosen m-digit pairs repeatedly, filling the last incomplete block with fixed digits. At the first differing block, rescale by the common prefix length. The linear sums differ by at least $2Dr^m$ before tails are added. Later digits change the difference by $O(r^m)$, and replacing $s(x - x')$ by $f(x) - f(x')$ introduces another $O(r^m)$ in the same coordinates by the mean value theorem. Choosing D larger than these constants proves separation at scale r^j. The fixed initial and final digits cost only a constant factor in the count. The strict margins in the gap and derivative bounds persist near (κ, λ) by continuity. A finite cover of the compact parameter rectangle gives common constants and a common lower threshold for j. Reducing c_ϵ handles the remaining finite lengths. This proves the finite-sum bound.

The lower bound. Fix the posterior x before two consecutive positive resets and their runs of j and $k = e + 2j$ continuations, in that order, where $e = \lceil H \rceil$. Write $u = h + \eta x$ and split the signed sum of the second word into its first $e + j$ digits B and its last j digits y, so $b_2 = B + r^{e+j} y$. If $z \in C_j$ is the first word's sum, direct substitution gives the final posterior as

$$\frac{r^k}{B + 1/h + r^{e+j}(y + f(z))}, \qquad \kappa = \eta 4^e, \quad \lambda = 1/u + \eta r^j / h.$$

These parameters lie in the rectangle above for every allowed x. The possible B are $\gg r^{e+j}$ apart. Retain a fixed fraction of them spaced far enough that the different sets of denominators have gaps $\gg r^{e+j}$, using the uniform bound on $|y + f(z)|$. Within each retained set use the separated digit pairs just proved. The denominators are bounded above and away from zero, so inversion preserves the order of these gaps. Thus there are at least $c_\epsilon 2^{e+(3-2\epsilon)j}$ final posteriors with gaps $\gg_\epsilon r^{2k}$. Each selected word pair has conditional probability $\gg 4^{-(e+3j)}$.

Take $j = \lceil (N - e + C_\epsilon)/(3 - 2\epsilon) \rceil$ with $0 < \epsilon < 1/2$ and C_ϵ large enough to leave at least $2L$ posteriors. At least L miss every one of L forecast centres by a fixed fraction of the gap, for each initial x. Integration gives

$$d_L^{\leftarrow} \gg_\epsilon L\, 4^{-(e+3j)} r^{4k} = 2^{N - 10e - 22j}.$$

Let $\eta \downarrow 0$ at fixed ϵ, then $\epsilon \downarrow 0$. Since $e = H + O(1)$ and $N = 2AH + o(1)$, this bounds $\limsup \log(d_L^{\leftarrow}) / \log \eta$ by $(19A + 4)/3$. The upper bound attained by the updating memory gives the opposite inequality for the lower limit. Since this memory is also a summary of the entire past, the two bounds prove the assertion for both kinds of memory.

B Next-symbol forecasts

For a stationary process $(X_t)_{t \in \mathbb{Z}}$ on a finite alphabet $\mathcal{A}$, let $\mu^{\rightarrow}$ be the law of the next-symbol forecast

$$Y = \Pr(X_0 = \cdot \mid X_{-1}, X_{-2}, \ldots)$$

on $\Delta_{\mathcal{A}}$. Write $R_L^{\rightarrow}$ for the infimum of $\mathbb{E} D_{\mathrm{KL}}(Y \| \Pr(X_0 = \cdot \mid S))$ over past summaries S with at most L values. Conditioning on the future defines $\mu^{\leftarrow}$ and $R_L^{\leftarrow}$. Distances are Euclidean.

Theorem B.1. *1.* $R_L^{\rightarrow} = \inf_{c_1, \ldots, c_L \in \Delta_{\mathcal{A}}} \int \min_k D_{\mathrm{KL}}(y \| c_k)\, d\mu^{\rightarrow}(y)$.

2. Suppose that for some $A, d > 0$ and $\rho_0 \geq 0$, $\mu^{\rightarrow}(B(y, \rho)) \leq A\rho^d$ for every y in the support and $\rho_0 \leq \rho \leq 1$. Then $R_L^{\rightarrow} \geq \frac{1}{16}(2AL)^{-2/d}$ for $1/(2A) \leq L \leq \rho_0^{-d}/(2A)$.

3. Suppose every coordinate of Y is at least $\delta > 0$ almost surely, and for some $C, D > 0$ and $\rho_0 \geq 0$ the support of $\mu^{\rightarrow}$ is covered by at most $C\rho^{-D}$ balls of radius ρ for $\rho_0 \leq \rho \leq 1$. Then $R_L^{\rightarrow} \leq 4\delta^{-1}(C/L)^{2/D}$ for $C \leq L \leq C\rho_0^{-D}$.

The same statements hold for forecasts of the next ℓ symbols, with $\mathcal{A}^\ell$ in place of $\mathcal{A}$.

Proof. For fixed centres, assigning Y to its cheapest forecast minimizes the loss. Equation (2) shows that replacing each centre by its cell mean can only improve the forecast. This proves part 1.

For part 2, take $\rho = \frac{1}{2}(2AL)^{-1/d}$. Each centre's ρ-ball meeting the support lies in a support-centred 2ρ-ball of mass at most $A(2\rho)^d = 1/(2L)$. At least half the mass is therefore farther than ρ from every centre. Pinsker's inequality gives $R_L^{\rightarrow} \geq \rho^2/4 = (2AL)^{-2/d}/16$.

For part 3, use the covering balls at $\rho = (C/L)^{1/D}$. Their cell means have coordinates at least δ and lie within 2ρ of every point of the cell. Hence $D_{\mathrm{KL}}(y \| c) \leq \sum_x (y_x - c_x)^2 / c_x \leq 4\rho^2/\delta$. For an ℓ-symbol block use the alphabet $\mathcal{A}^\ell$. □

A covering bound controls the required number of states; a matching mass bound gives the same lower exponent. The next example explains the positive lower bound required for forecast probabilities.

Remark B.2 (Near the boundary)**.** *The positive lower bound on forecast probabilities in Theorem B.1 matters for a uniform upper estimate. Consider equally weighted Bernoulli forecasts k/N^2 for $0 \le k \le N$. Their covering number at scale ρ is $O(\rho^{-1/2})$ uniformly in N, by taking the smaller of $N+1$ and $O(1+1/(N\rho))$. The curvature of negative Bernoulli entropy on $(0, 1/N]$ is at least N. For a cell of m grid points, the sum of squared distances from the cell mean is at least $(m-1)/(2N^4)$ by the pairwise variance identity. Summing over the cells gives a loss of at least $(N+1-L)/(4(N+1)N^3)$. With $L = \lfloor N/2 \rfloor$, this gives $L^4 R_L \gg N$, despite the uniform covering bound with $D = 1/2$.*

References

[1] J. W. Carlyle and A. Paz, Realizations by stochastic finite automata, *J. Comput. Syst. Sci.* **5** (1971), 26–40.

[2] M. Hochman, On self-similar sets with overlaps and inverse theorems for entropy, *Ann. of Math.* (2) **180** (2014), 773–822. doi:10.4007/annals.2014.180.2.7.

[3] M. Hochman and B. Solomyak, On the dimension of Furstenberg measure for $\mathrm{SL}_2(\mathbb{R})$ random matrix products, *Invent. Math.* **210** (2017), 815–875.

[4] D.-J. Feng and H. Hu, Dimension theory of iterated function systems, *Comm. Pure Appl. Math.* **62** (2009), 1435–1500. doi:10.1002/cpa.20276.

[5] M. Hochman and B. Solomyak, Clarification to the dimension theorem for $\mathrm{SL}_2(\mathbb{R})$ random matrix products, authors' note (2018).

[6] D. Hsu, S. M. Kakade and T. Zhang, A spectral algorithm for learning hidden Markov models, *J. Comput. Syst. Sci.* **78** (2012), 1460–1480. arXiv:0811.4413.

[7] J. Berstel and C. Reutenauer, Zeta functions of formal languages, *Trans. Amer. Math. Soc.* **321** (1990), 533–546.

[8] R. Bowen and O. E. Lanford III, Zeta functions of restrictions of the shift transformation, in *Global Analysis*, Proc. Sympos. Pure Math. **14**, Amer. Math. Soc., 1970, 43–49.

[9] T. Y. Chien and S. Waldron, A characterization of projective unitary equivalence of finite frames and applications, *SIAM J. Discrete Math.* **30** (2016), 976–994.

[10] J. P. Crutchfield, C. J. Ellison and J. R. Mahoney, Time's barbed arrow: irreversibility, crypticity, and stored information, *Phys. Rev. Lett.* **103** (2009), 094101.

[11] C. J. Ellison, J. R. Mahoney, R. G. James, J. P. Crutchfield and J. Reichardt, Information symmetries in irreversible processes, *Chaos* **21** (2011), 037107.

[12] P. Feng and C. Meng, Recursive memory for prediction and generation, arXiv:2611.13760 (2026).

[13] E. Gassiat, A. Cleynen and S. Robin, Inference in finite state space non parametric hidden Markov models and applications, *Stat. Comput.* **26** (2016), 61–71.

[14] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Lecture Notes in Math. **1730**, Springer, 2000.

[15] S. Graf and H. Luschgy, The quantization dimension of self-similar probabilities, *Math. Nachr.* **241** (2002), 103–109.

[16] M. Kesseböhmer, A. Niemann and S. Zhu, Quantization dimensions of compactly supported probability measures via Rényi dimensions, *Trans. Amer. Math. Soc.* **376** (2023), 4661–4678. doi:10.1090/tran/8863.

[17] F. P. Kelly, *Reversibility and Stochastic Networks*, Wiley, 1979.

[18] L. J. Lindsay and R. D. Mauldin, Quantization dimension for conformal iterated function systems, *Nonlinearity* **15** (2002), 189–199.

[19] R. Loewy, Principal minors and diagonal similarity of matrices, *Linear Algebra Appl.* **78** (1986), 23–64.

[20] S. E. Marzen and J. P. Crutchfield, Predictive rate-distortion for infinite-order Markov processes, *J. Stat. Phys.* **163** (2016), 1312–1338.

[21] S. E. Marzen and J. P. Crutchfield, Nearly maximally predictive features and their dimensions, *Phys. Rev. E* **95** (2017), 051301(R).

[22] P. Mattila, *Geometry of Sets and Measures in Euclidean Spaces*, Cambridge Studies in Advanced Mathematics **44**, Cambridge University Press, 1995.

[23] V. Papadopoulos, J. Wenger and C. Hongler, Arrows of time for large language models, arXiv:2401.17505.

[24] A. Rapaport, Dimension of self-conformal measures associated to an exponentially separated analytic IFS on $\mathbb{R}$, preprint, arXiv:2412.16753v2.

[25] M. K. Roychowdhury, Quantization dimension function and ergodic measure with bounded distortion, *Bull. Polish Acad. Sci. Math.* **57** (2009), 251–262. Publisher's PDF.

[26] K. S. Sibirskii, Algebraic invariants of a system of matrices, *Siberian Math. J.* **9** (1968), 115–124.

[27] B. Solomyak, On nonlinear iterated function systems with overlaps, *J. Fractal Geom.* **11** (2024), 373–387.

[28] B. Solomyak and Y. Takahashi, Diophantine property of matrices and attractors of projective iterated function systems in $\mathbb{RP}^1$, *Int. Math. Res. Not. IMRN* (2021), 12639–12669.

[29] C. Yang, A. J. P. Garner, F. Liu, N. Tischler, J. Thompson, M.-H. Yung, M. Gu and O. Dahlsten, Provably superior accuracy in quantum stochastic modeling, *Phys. Rev. A* **108** (2023), 022411.