hexagon:2611.01320v1

Classical and Quantum Memory for Conditional Future Generation

Zhongpai Gao
United Imaging Intelligence, Boston, MA, USA

Abstract

A finite memory observes a hidden Markov source one symbol at a time and then generates its conditional future. We measure the Kullback–Leibler divergence between the laws of the entire future after averaging the device's private randomness. Forecasts are linear in a space of dimension $r_D = D$ for D classical states and $r_D = D^2$ for quantum dimension D. We construct positive updates that track functions of the hidden posterior. For every strictly positive source of word-probability rank $\rho \geq 2$, the resulting error is $O(\exp(-cr_D^{1/(\rho-1)}))$. For invertible sources whose rank equals their hidden-state count, a unified converse uses matrices of reciprocal likelihood pairings. It gives matching lower bounds when the posterior and likelihood sets both contain open regions. This class includes a nonempty open set of n-state, n-symbol sources for each n. It also gives matching bounds for every positive invertible non-iid two-state source. In these cases, accuracy ε requires order $\log^{n-1}(1/\varepsilon)$ classical states and its square root in quantum dimension. Two boundary examples show why the assumptions matter: an invertible three-state source on a posterior curve has a faster upper bound, while a singular two-state source has error $\exp[-\Theta(r_D \log(r_D + 1))]$. All persistent observation and generation memory is counted. Constants depend on the known source; finite precision and running time are not addressed.

Index terms: Hidden Markov models, quantum memory, approximation, relative entropy.

1 Introduction

A hidden Markov source emits symbols while its internal state stays hidden. Suppose an observer has watched the symbols up to now and must then continue the sequence: it must generate a future with the correct conditional distribution, given what it has seen. With unlimited memory this is easy. The observer keeps the conditional distribution of the hidden state, updates it after every symbol, and at the end samples a hidden state from it and runs the source. This distribution, the *posterior*, usually takes infinitely many values, so a finite memory cannot keep it exactly. How much memory is needed to generate the future within a given error, and does a quantum memory need less?

We study memories that read one symbol at a time. A classical memory has D states and makes a random transition that depends on the symbol. A quantum memory is a register of dimension D and undergoes a quantum channel that depends on the symbol. When reading stops, the device generates the future from the same memory. Its error is the Kullback–Leibler divergence from the true conditional law of the entire future to the law of the generated future, averaged over the observed pasts. The generated law is taken after averaging over the device's own randomness: the device is judged by what it produces, not by each of its internal states separately.

1.1 The main idea

The law that a device generates is a linear function of its average memory state given the past. This average state is a probability vector with D entries or a $D \times D$ density matrix, and reading a symbol

changes it by a linear map. A device therefore carries

$$r_D = D \quad \text{(classical)} \qquad \text{or} \qquad r_D = D^2 \quad \text{(quantum)}$$

linear coordinates, updates them linearly, and reads its forecasts off linearly. Normalization removes one independent coordinate, but the linear factorization can still have rank r_D. The source updates its posterior by a different kind of rule, a ratio of linear functions. The memory question thus becomes a question of approximation: how well can this nonlinear recursion be imitated by a linear recursion in r_D variables? Both our upper and lower bounds answer this question, and the type of memory enters only through r_D. That a quantum system of dimension D offers a D^2-dimensional linear representation is well known [16, 17, 18]; the work here is to obtain valid updates, normalized conditional laws and matching bounds for the divergence.

1.2 Results

Throughout, every symbol matrix of the source is strictly positive. Let n be the number of hidden states and $\rho_S \leq n$ the rank of the matrix of word probabilities; rank one means iid output.

An upper bound for every source. Theorem 3.1 constructs, for every source with $\rho_S \geq 2$ and every $D \geq n$, a device with

$$\text{error} \;\leq\; Ce^{-c\, r_D^{1/m}}, \qquad m = \rho_S - 1.$$

The device stores the monomials of degree at most K in m coordinates of the posterior. There are about K^m of them, and their update is linear up to an error of size e^{-cK}.

A lower bound from test points. Theorem 4.4 applies when the symbol matrices are also invertible and $\rho_S = n$. Choose M hidden-state distributions p_i that long pasts can produce, and M columns h_j of likelihoods that long continuations can produce. Form the $M \times M$ matrix of reciprocals $1/\langle p_i, h_j \rangle$. If its least singular value is s, every device with $r_D < M$ has error at least $c\,(s/M)^\kappa$. The constants depend only on the source. A lower bound is thereby reduced to choosing well-conditioned test points in two sets.

Two consequences. First, suppose the hidden-state distributions produced by long pasts fill an open region of the simplex, and the same holds for the likelihood columns. Then test points can be placed on a grid, and Theorem 5.1 gives

$$\text{error} \;\geq\; e^{-C\, r_D^{1/(n-1)}},$$

which matches the upper bound. This class contains a nonempty open set; for every n there is an explicit example with n hidden states and n symbols (Corollary 5.5). Second, for two hidden states the lower bound holds for every invertible source that is not iid, whatever the shape of these sets (Theorem 6.1).

With $L_\varepsilon = \log(1/\varepsilon)$, the memory needed for error ε is:

Source	Classical states	Quantum dimension
Any positive source of rank ρ_S (upper bound)	$O(L_\varepsilon^{\rho_S - 1})$	$O(L_\varepsilon^{(\rho_S-1)/2})$
Invertible, n states, both sets fill open regions	$\Theta(L_\varepsilon^{n-1})$	$\Theta(L_\varepsilon^{(n-1)/2})$
Invertible, two states, not iid	$\Theta(L_\varepsilon)$	$\Theta(\sqrt{L_\varepsilon})$

In the matched cases the quantum dimension is the square root of the classical state count. As $\varepsilon \downarrow 0$, the leading qubit cost is half the leading bit cost; this is not a finite-budget replacement

rule. The mechanism is the D^2 versus D linear dimension. Our quantum devices measure and re-prepare their register at every block update; they use the larger state space, not interference between computations.

A limit. Appendix A gives a source with three hidden states, rank three and invertible matrices whose posteriors lie on a curve. It admits error $O(e^{-cr_D})$, a faster upper bound than $O(e^{-cr_D^{1/2}})$ for rank three; no matching lower bound is proved here. The rank alone therefore does not determine the rate, and some condition such as the open-region hypothesis is needed for the lower bound. For three or more hidden states, sources whose posteriors do not fill an open region are not classified. There is also a distinct boundary at singular matrices: Appendix B proves the faster order $\exp[-\Theta(r_D \log(r_D + 1))]$ for one positive two-state source. It explains why invertibility cannot simply be omitted.

1.3 How the lower bound is proved

Fix some words that may be observed next, some words that may be observed after them, and some words that may be generated after that. The true probabilities of the generated words form a table, indexed by the first group of words on one side and the other two groups on the other. The device's probabilities form a second table, and it has rank at most r_D because all its forecasts pass through its memory after the first group of words (Lemma 4.2). For an invertible source, a fixed combination of the forecasts of the next $n - 1$ symbols equals the reciprocal of the probability of the word just observed (Lemma 4.3). The same combination of the device's table still has rank at most r_D. The true table becomes the matrix of reciprocals above, and its distance from matrices of rank r_D is at least its least singular value.

In suitable coordinates the reciprocal $1/\langle p, h\rangle$ is the kernel $1/(1 - y \cdot \zeta)$. On a product grid this kernel is bounded below by a product of Vandermonde matrices, which gives the first consequence. For two hidden states it is a Cauchy matrix, whose inverse is explicit, which gives the second.

1.4 Earlier work

Prediction and generation can have different memory costs [4, 15], and the number of features needed for prediction reflects the geometry of the posteriors [5]. Quantum devices that apply a channel for each supplied symbol are established [1, 2]. Earlier quantum advantages concern entropy at finite precision [3], exact simulation of three classical states by one qubit [6], continuous-variable sequence models [7], and approximation measured by divergence rates, fidelity, finite horizons or selected events [8, 13, 14, 9, 10].

Feng and Meng consider the divergence of the entire future when a deterministic observer is followed by a randomized generator [11, Section 4.5]. We allow randomized observers and prove matching state counts that are polynomial in $\log(1/\varepsilon)$. Estimates for filters with a finite window [20] give exponential accuracy in the window length, but storing a window of length k costs $|\mathcal{A}|^k$ states. Our problem differs from the approximation of weighted automata as linear systems [19], where positivity of the updates is not required, and from language recognition with a strict cutpoint [12]. The author's paper on time reversal [22] scores each memory state separately instead of averaging the device's randomness first; the rates there are power laws.

1.5 Organization

Section 2 states the problem and collects the tools for building devices. Section 3 proves the upper bound. Section 4 proves the general lower bound, Section 5 applies it to sources whose posteriors

fill an open region, and Section 6 to two hidden states. Appendices A and B prove the curve and singular boundary examples. Finite-budget certificates, parameter neighborhoods and superseded constructions are retained separately; no main result depends on them.

2 The problem

All logarithms are natural. A classical memory has D states and a quantum memory is a register of dimension D. We use $\diamond \in \{\mathrm{cl}, \mathrm{q}\}$ to select the memory type and write

$$r_D = \begin{cases} D & \diamond = \mathrm{cl}, \\ D^2 & \diamond = \mathrm{q}. \end{cases} \tag{1}$$

These are the dimensions of the real linear spaces containing a probability row or a Hermitian matrix. Normalization leaves $D-1$ or D^2-1 free coordinates, but linear forecast matrices can have rank D or D^2. We use r_D for that rank bound.

2.1 The source

A hidden Markov source has n hidden states and a finite alphabet $\mathcal{A}$. It is specified by $n \times n$ matrices $T_a = T^a$, one for each symbol, whose entries are positive and whose sum is row-stochastic. The entry $T_a(i,j)$ is the probability of emitting a and moving from hidden state i to j. The matrices are known to the device. Their sum has a unique stationary row π, and we use the stationary source: if Z_t is the hidden state after the symbol X_t, then $Z_t \sim \pi$ and

$$\Pr(X_{t+1} = a, Z_{t+1} = j \mid X_{\le t}, Z_{\le t}) = T_a(Z_t, j).$$

For an observed past H, let p_H be the conditional distribution of the current hidden state, written as a row. We call it the *posterior*. For any row p of hidden-state probabilities, write P_p for the law of the future when the hidden state starts with distribution p. Then

$$P_p = \sum_{i=1}^{n} p_i P_{e_i}, \qquad \mathcal{F}_a(p) = \frac{pT_a}{pT_a\mathbf{1}}, \tag{2}$$

where e_i is the ith point mass and $\mathcal{F}_a$ is the update of the posterior when a is observed. For a word $w = a_1 \cdots a_k$ put $T_w = T_{a_1} \cdots T_{a_k}$ and $\mathcal{F}_w(p) = pT_w/(pT_w\mathbf{1})$. Then

$$\Pr(X_{t+1:t+|w|} = w \mid X_{\le t}) = p_{X_{\le t}} T_w \mathbf{1}, \qquad p_{Hw} = \mathcal{F}_w(p_H). \tag{3}$$

These identities hold first for finite histories; martingale convergence gives them for the full past. If $\varpi = \min_{a,i}(T_a\mathbf{1})_i$, every conditional word probability is at least $\varpi^{|w|} > 0$.

The conditional law of the future given the past H is P_{p_H}. It depends on H only through the posterior, but the posterior usually takes infinitely many values.

2.2 The device and its error

The device reads the symbols of the source one at a time. When it reads a, a classical device replaces its current state s by a random state drawn from row s of a fixed stochastic matrix K_a. A quantum device applies a fixed trace-preserving completely positive (CPTP) map Φ_a to its density matrix on $\mathbb{C}^D$. The symbol comes from the source; it is not obtained by measuring the memory. A quantum update must be a valid channel on every input.

At a chosen time, the *cut*, reading stops and the device must produce the future itself. It may first apply one fixed channel to its memory. It then emits symbols through nonnegative matrices, one for each symbol, whose sum is stochastic, or through completely positive maps whose sum is trace preserving. Reading and generating use the same memory, and everything the device keeps from one step to the next counts in D.

Let Q_H be the law of the generated future given the observed past H, after averaging over the device's own randomness. The error is

$$\mathcal{R} = \mathbb{E}_H D_{\mathrm{KL}}(P_{p_H} \| Q_H), \qquad D_{\mathrm{KL}}(P\|Q) = \sup_{k\geq 1} D_{\mathrm{KL}}(P_{1:k}\|Q_{1:k}), \tag{4}$$

where $P_{1:k}$ is the law of the first k symbols and $D_{\mathrm{KL}}(P_{1:k}\|Q_{1:k}) = \sum_w P(w)\log(P(w)/Q(w))$. This is the divergence between the laws of the entire future; it is not divided by the length of the future. We call $\mathcal{R}$ the *risk*.

The averaging matters. By convexity, if S is the memory state,

$$\mathbb{E}_H D_{\mathrm{KL}}\big(P_{p_H} \| \mathbb{E}[Q_S \mid H]\big) \leq \mathbb{E}_{H,S} D_{\mathrm{KL}}(P_{p_H}\|Q_S).$$

The right-hand side scores each memory state separately and can need much more memory. We use the left-hand side: the device is judged by the distribution of what it generates after each past.

Admitted devices. The device is stationary and causal. For a classical device this means a jointly stationary law of $(X_t, S_t)_{t\in\mathbb{Z}}$ with the source as its first marginal, such that

$$\Pr(S_t = s' \mid X_{\leq t}, S_{<t}) = K_{X_t}(S_{t-1}, s'), \tag{5}$$

$$S_{\leq t} \perp X_{>t} \mid X_{\leq t}. \tag{6}$$

The second condition says that the memory carries no information about future symbols beyond what the observed past carries. No other condition is imposed in the lower bounds. For a quantum device, let $\mathcal{H} = \mathcal{A}^{\mathbb{Z}_{\leq 0}}$ be the space of observed pasts. One measurable function ρ on $\mathcal{H}$ assigns a density matrix to each past, the register at time t is in the state $\rho(X_{\leq t})$, and $\rho(Ha) = \Phi_a(\rho(H))$ for almost every H and every symbol a. The generated law is linear in this density matrix, as the classical generated law is linear in the conditional distribution ξ_H of the memory state.

The device may use fresh independent randomness and temporary workspace at each step. Anything it retains, such as a record of earlier outcomes, a random seed or a system entangled with the register, is part of the memory and counts in D. In particular a random choice between different update rules is not free: the device would have to remember which rule it chose. The fixed description of the update rules is not counted. The maps may have arbitrary real coefficients; we do not study finite precision or total hardware cost.

For a source S, let $G_D^{\mathrm{cl}}(S)$ and $G_D^{\mathrm{q}}(S)$ be the infima of the risk over classical and quantum devices of size D, and write $G_D^{\diamond}(S)$ for either. Let $M_{\mathrm{cl}}(\varepsilon; S)$ and $M_{\mathrm{q}}(\varepsilon; S)$ be the least sizes of devices with risk at most ε.

Lemma 2.1. *The conditional distribution of the classical memory state, $\xi_H = \Pr(S_t \in \cdot \mid X_{\leq t} = H)$, satisfies $\xi_{Ha} = \xi_H K_a$. For a word $w = a_1 \cdots a_k$ write $K_w = K_{a_1} \cdots K_{a_k}$ and $\Phi_w = \Phi_{a_k} \circ \cdots \circ \Phi_{a_1}$. Then $\xi_{Hw} = \xi_H K_w$ and $\rho(Hw) = \Phi_w(\rho(H))$ for every word, on one set of pasts of full measure.*

Proof. Condition (6) at time $t-1$ gives $\Pr(S_{t-1} = s \mid X_{\leq t}) = \Pr(S_{t-1} = s \mid X_{\leq t-1})$. Conditioning (5) on $X_{\leq t}$ therefore gives $\xi_{Ha}(s') = \sum_s \xi_H(s) K_a(s, s')$ on $\{X_t = a\}$. The quantum identity is part of the model. Iteration gives the word identities. There are countably many words and each has positive conditional probability, so one set of full measure serves all of them. □

2.3 From hidden-state error to entire-future error

A device that starts the exact source from a slightly wrong hidden-state distribution makes a small error on the entire future.

Lemma 2.2. *For any n-state source and probability rows p, q with $q_i > 0$,*

$$D_{\mathrm{KL}}(P_p \| P_q) \le D_{\mathrm{KL}}(p\|q) \le \sum_i \frac{(p_i - q_i)^2}{q_i}. \tag{7}$$

Proof. At every finite horizon the future is obtained from the initial hidden state by a fixed random map. Such a map cannot increase the divergence, and the divergence is at most the chi-square distance; taking the supremum over horizons proves (7). □

2.4 Constructing stationary memories

A device in our model has no starting time. The next lemma shows how to define such a device from update rules: it constructs the memory at every time from the remote past. It is used only to build the devices of the upper bounds, and places no restriction on the devices covered by the lower bounds. In the lemma the update applied at time t may depend on the observed past; in each application we check separately that the current symbol and the stored memory determine it.

Lemma 2.3 (Memories defined from the remote past)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be stationary. Suppose a finite-dimensional sequence of local maps is measurable from $X_{\le t}$ and is shift-covariant: moving the input sequence by one time unit moves the sequence of maps by the same unit.*

1. *For classical kernels K_t, suppose a fixed probability row ν and $\eta > 0$ satisfy $\mathsf{K}_t(i,j) \ge \eta\nu_j$ for all t, i, j. There is an actual jointly stationary causal label process with these conditional transitions, using fresh private coins independent of the entire source. Its conditional mean rows satisfy $z_t = z_{t-1}\mathsf{K}_t$.*

2. *For CPTP maps $\mathcal{K}_t$, suppose $\|\mathcal{K}_t(\rho - \sigma)\|_1 \le q\|\rho - \sigma\|_1$ for all density matrices and a fixed $0 \le q < 1$. Composing from the remote past gives a unique past-measurable stationary density matrix $\sigma_t = \mathcal{K}_t(\sigma_{t-1})$, independent of the starting state.*

Proof. For the first assertion, if $\eta = 1$ prepare ν at each step. Otherwise write $\mathsf{K}_t = \eta\mathbf{1}\nu + (1-\eta)\widetilde{\mathsf{K}}_t$. At every time take an independent Bernoulli coin of parameter η, a fresh sample from ν, and a uniform variable for sampling a row of $\widetilde{\mathsf{K}}_t$. All these variables are independent of the entire source. A successful coin replaces the label by its fresh sample; otherwise the old label is updated with $\widetilde{\mathsf{K}}_t$. This realizes exactly K_t.

Among the last k times, the probability of no replacement is $(1-\eta)^k$. Thus a most recent replacement before any prescribed t exists almost surely. Start there and apply the finitely many subsequent updates. This defines the label simultaneously for every integer t, by a countable intersection of probability-one events. All earlier initial labels give the same result once a replacement has occurred. The replacement time is used to define the process, not retained by the online device.

The label up to time t is a function of the observed past and private variables up to t. Its conditional transitions are the stated kernels, since the next coins are fresh. Given the observed past, these private variables are independent of future source symbols, proving (6) and the mean-row recursion. The construction commutes with time shifts; the source and independent coin sequences have stationary joint law. This proves joint stationarity.

For the second assertion, let $\sigma_t^{[k]}$ be the result of the last k maps applied to any starting density matrix. If the start is moved farther back, the common last k maps give

$$\|\sigma_t^{[k]} - \sigma_t^{[\ell]}\|_1 \le 2q^k, \qquad \ell \ge k.$$

The sequence is uniformly Cauchy. Its limit is a density matrix, is independent of the start, and is measurable from the past as a limit of measurable finite compositions. Continuity gives the update identity; shift covariance of the compositions passes to the limit and proves stationarity. No quantum reference or private state outside the local register is used. □

The lemma also applies when updates occur only at the ends of blocks of a fixed length: apply it to the process observed at those times, for each alignment of the blocks, and then choose the alignment uniformly at random. It does not assert contraction of a register that also stores the position inside a block, since different positions never merge.

2.5 Implementing affine updates

To retain functions $r_j(x)$ of an unknown posterior, the device must update their means from the observed symbols alone. We implement affine updates of the form

$$r \longmapsto b + \mathsf{M}r, \qquad a_0 := \sum_j |b_j|, \quad a_1 := \sum_{j,\ell} |\mathsf{M}_{j\ell}|, \quad a_0 + a_1 < 1. \tag{8}$$

For a power-of-two dimension d, use the $M = d^2 - 1$ nonidentity Hermitian Pauli strings P_j (tensor products of I, X, Y, Z), with P_1 a single Pauli Z. They have eigenvalues ± 1 and $\operatorname{tr}(P_j P_\ell) = d\,\mathbf{1}_{j=\ell}$. If $\sum_j |r_j(x)| \le 1$, then $\rho_x = (I + \sum_j r_j(x) P_j)/d$ is a density matrix with $\operatorname{tr}(P_j \rho_x) = r_j(x)$.

Lemma 2.4. *The affine update* (8) *is realized by a CPTP map on all d-dimensional inputs. On trace-zero differences it contracts trace norm by at most a_1. For every power of two N, the same coordinate update with $M = N - 1$ is realized by an N-state stochastic map; its mean-feature errors contract in coefficient ℓ^1 norm by at most a_1.*

Proof. On unnormalized inputs X, let

$$\mathcal{D}_0(X) = \operatorname{tr}(X) I/d, \qquad \mathcal{P}_j^s(X) = \operatorname{tr}(X)(I + sP_j)/d,$$
$$\mathcal{C}_{j\ell}^s(X) = \operatorname{tr}(X) I/d + s \operatorname{tr}(P_\ell X) P_j/d, \qquad s = \pm 1.$$

The last channel measures P_ℓ and prepares $(I + svP_j)/d$ after outcome v. The positive mixture

$$\Phi = (1 - a_0 - a_1)\mathcal{D}_0 + \sum_j |b_j| \mathcal{P}_j^{\operatorname{sign} b_j} + \sum_{j,\ell} |\mathsf{M}_{j\ell}| \mathcal{C}_{j\ell}^{\operatorname{sign} \mathsf{M}_{j\ell}} \tag{9}$$

is therefore CPTP and has coordinates $b + \mathsf{M}r$. Zero terms are omitted. Preparations vanish on trace-zero differences, while every $\mathcal{C}_{j\ell}^s$ contracts their trace norm by at most one.

For the classical construction write s, j as $\log_2 N$-bit strings and set $t_{s,j} = (-1)^{s \cdot j}$, omitting $j = 0$. These rows form a matrix of signs with orthogonal columns, usually called a Walsh matrix. Thus $\sum_s t_s = 0$ and $\sum_s t_s t_s^{\mathsf{T}} = NI$. Define

$$\mathsf{K}(s, v) = \frac{1 + t_v^{\mathsf{T}}(b + \mathsf{M}t_s)}{N}. \tag{10}$$

Its entries are at least $(1 - a_0 - a_1)/N$, each row sums to one, and its expected destination feature is $b + \mathsf{M}t_s$. Since $\|\mathsf{M}\|_{1\to 1} \le a_1$, its mean features have the claimed error bound. A reference feature $r(x)$ is prepared with probabilities $(1 + t_s^{\mathsf{T}} r(x))/N$ whenever $\|r(x)\|_1 \le 1$. □

2.6 Tracking functions of the posterior

The devices of the upper bounds store finitely many functions of the posterior, such as its powers. We call them *features*. The next lemma bounds the error of a device whose stored features follow an approximately correct affine recursion, and which at the cut reads a hidden-state distribution off them and runs the source exactly.

Lemma 2.5 (Updating features and reading the future)**.** *Suppose bounded reference features z_t and actual mean features $\widehat{z}_t$ obey, at successive updates,*

$$z_t = b_t + A_t z_{t-1} + e_t, \qquad \widehat{z}_t = b_t + A_t \widehat{z}_{t-1}, \qquad \|A_t\|_{1\to 1} \le \kappa < 1, \quad \|e_t\|_1 \le \tau.$$

If both recursions extend arbitrarily far into the past with uniformly bounded values, their difference is at most $\tau/(1-\kappa)$. The same bound holds after an exact preparation. Suppose a valid affine cut readout $a + L\widehat{z}_t = \widehat{p}_t$ starts the exact source generator on the same register. If $\|L\|_{1\to 1} \le B$, $\widehat{p}_{t,i} \ge \mu > 0$, and

$$\|p_t^* - (a + Lz_t)\|_1 \le \epsilon, \qquad P_{p_t^*} = P_{p_{H_t}},$$

where p_t^* is a hidden probability row, then

$$D_{\mathrm{KL}}(P_{p_{H_t}} \| P_{\widehat{p}_t}) \le \frac{1}{\mu}\left(\epsilon + \frac{B\tau}{1-\kappa}\right)^2. \tag{11}$$

The lower margin is required on reached mean readouts, not on every physical input. The readout itself must be positive and normalized on every input.

Proof. After k updates the feature error is at most $\kappa^k E_0 + \tau \sum_{j=0}^{k-1} \kappa^j$. Bounded remote-past starts give the claim as $k \to \infty$; an exact preparation has $E_0 = 0$. The readout error is at most $\epsilon + B\tau/(1-\kappa)$ in ℓ^1. Lemma 2.2 and $\sum_i v_i^2 \le \|v\|_1^2$ prove (11) for the entire future. □

Lemma 2.4 implements the affine part when its coefficient mass is small. For power-of-two register sizes, M features fit N classical states when $M \le N - 1$, or quantum dimension d when $M \le d^2 - 1$. Unused states or dimensions are filled with any valid stochastic or CPTP rule. If the device is a mixture of several components whose priors have the same true future law, the lemma is applied to each component and convexity of the divergence gives the bound for the mixture. The bound does not grow with the length of the future.

3 An upper bound for every positive source

Let π be the stationary hidden law. The *word-probability rank* ρ_S of the source is the rank of the infinite matrix $[\pi T_u T_v \mathbf{1}]_{u,v}$ whose rows and columns are indexed by all finite words, including the empty word. Its entries are the probabilities of the words uv. The rank is at most the number n of hidden states, and it is smaller when some hidden directions never affect the observations.

Theorem 3.1. *Let S have $n \ge 2$ hidden states and a finite alphabet $\mathcal{A}$, with every symbol matrix strictly positive and their sum stochastic. If $m = \rho_S - 1 \ge 1$, there are source-dependent $C_S, c_S > 0$ such that, for every $D \ge n$,*

$$G_D^{\diamond}(S) \le C_S e^{-c_S r_D^{1/m}}, \qquad \diamond \in \{\mathrm{cl}, \mathrm{q}\}.$$

If $\rho_S = 1$, the observed process is iid and one state or dimension is exact. In particular every positive n-state source admits the exponents $D^{1/(n-1)}$ and $D^{2/(n-1)}$. The devices are causal and stationary, apply one fixed update for each observed symbol, and generate on the same register at every cut.

Idea of the proof. A memory cannot store the posterior exactly, and it cannot update a stored approximation exactly either: the posterior update is a ratio of linear functions, while the average state of a memory can only be updated linearly. The device therefore stores the monomials of degree at most K in m coordinates of the posterior. After one more observation, each monomial of the new coordinates is a power series in the old ones, and dropping the terms of degree above K leaves a linear update of the stored monomials. There are about K^m monomials, so they fit a memory with $r_D \approx K^m$, and the dropped terms have size e^{-cK}. This gives the rate.

Three points need care. First, the power series must have small coefficients for the linear update to be a valid stochastic map or quantum channel. This holds after a block of b_0 symbols, because the posterior then depends only weakly on where it started. The device therefore updates its monomials once per block and keeps the symbols of the current block in a buffer. Second, a stationary device has no clock, so the position inside the block is itself stored, as the length of the buffer. Third, the device must generate from what it has stored: at the cut it measures an approximate hidden state and then runs the source exactly. The six steps of the proof take these points in order.

The buffer and block labels multiply the memory size by a factor that depends only on the source. With block length $b_0 = b_0(S) \geq n$, the labels are

$$\mathcal{U} = \mathcal{A}^{b_0}, \quad \mathcal{W} = \bigcup_{j=b_0}^{2b_0-1} \mathcal{A}^j, \quad H_S = |\mathcal{U}||\mathcal{W}| = |\mathcal{A}|^{b_0} \sum_{j=b_0}^{2b_0-1} |\mathcal{A}|^j. \tag{12}$$

Here $u \in \mathcal{U}$ is the last completed block and $w \in \mathcal{W}$ is the buffer. A register of N states or dimension d for the monomials therefore costs $H_S N$ classical states or quantum dimension $H_S d$ in total. Generation uses n of these states or vectors again; no extra memory is added for it.

Proof. Write Δ for the hidden simplex and $\mathcal{F}_w(x) = xT_w/(xT_w\mathbf{1})$, with matrices in chronological order. *1. Keep only directions that affect the future.* Define spaces of real rows, including the empty word, by

$$\mathcal{V} = \operatorname{span}\{\pi T_w : w \in \mathcal{A}^*\}, \qquad \mathcal{Z} = \{\zeta : \zeta T_w \mathbf{1} = 0 \text{ for every } w\}, \qquad I_0 = \mathcal{V} \cap \mathcal{Z}.$$

Both $\mathcal{V}$ and $\mathcal{Z}$ are right-invariant under every T_a. The map $\zeta \mapsto (\zeta T_v \mathbf{1})_v$ on $\mathcal{V}$ has kernel I_0 and the row span of the word-probability matrix as its image, so

$$\rho_S = \dim \mathcal{V} - \dim I_0, \qquad V_0 = \{\zeta \in \mathcal{V} : \zeta \mathbf{1} = 0\}, \qquad \dim(V_0/I_0) = \rho_S - 1.$$

The empty word gives $I_0 \subseteq V_0$; $\pi\mathbf{1} = 1$ gives $\dim V_0 = \dim \mathcal{V} - 1$. If $\rho_S = 1$, the quotient is generated by π and $\pi T_a \equiv s_a \pi \pmod{I_0}$, where $s_a = \pi T_a \mathbf{1}$. Right invariance gives $\pi T_w \mathbf{1} = \prod_j s_{a_j}$. Thus the output is iid and has a one-state exact generator. Henceforth $m = \rho_S - 1 \geq 1$. Choose linear maps

$$\Pi : \mathbb{R}^n \longrightarrow \mathbb{R}^m, \quad \Lambda : \mathbb{R}^m \longrightarrow V_0, \quad \Pi\Lambda = I, \quad \ker \Pi \cap V_0 = I_0.$$

Put $L_\Pi = \|\Pi\|_{1\to 1}$, $\lambda_j = \Lambda e_j$ and $L_\Lambda = \max_j \|\lambda_j\|_1 = \|\Lambda\|_{1\to 1}$, so $L_\Pi L_\Lambda \geq 1$. Both word-generated spans stabilize within dimension n, so the maps are obtained by finite linear algebra from the known source. Finite-past posteriors are normalized πT_w; their martingale limits, the full-past posteriors, remain in $\mathcal{V}$. If $x, y \in \mathcal{V} \cap \Delta$ and $x - y \in I_0$, then

$$P_x = P_y, \qquad \mathcal{F}_w(x) - \mathcal{F}_w(y) = \frac{(x-y)T_w}{xT_w\mathbf{1}} \in I_0. \tag{13}$$

The denominators agree because all future word probabilities agree; right invariance gives the second identity.

2. Forget earlier hidden states uniformly. Set

$$\vartheta = \min_{a,i,i',j} \frac{T_a(i,j)}{T_a(i',j)}, \quad \beta = \min\{\vartheta^2, 1/2\}, \quad \mu = \min_{a,i,j} \frac{T_a(i,j)}{(T_a\mathbf{1})_i}, \quad \varepsilon_\ell = 2(1-\beta)^\ell.$$

For a word $w = a_1 \cdots a_\ell$, write $h^{(k)} = T_{a_{k+1}} \cdots T_{a_\ell}\mathbf{1}$ and

$$S_k = \operatorname{diag}(h^{(k-1)})^{-1} T_{a_k} \operatorname{diag}(h^{(k)}).$$

These matrices are stochastic. The definition of ϑ gives $T_{a_k}(i,j) \geq \vartheta T_{a_k}(i',j)$ and, after summation, $h^{(k-1)}_{i'}/h^{(k-1)}_i \geq \vartheta$. Hence every row of S_k dominates $\beta S_k(1,\cdot)$: zero-mass row differences contract in ℓ^1 by at most $1-\beta$. The rows of $S_1 \cdots S_\ell = \operatorname{diag}(T_w\mathbf{1})^{-1} T_w$ are $\mathcal{F}_w(e_i)$, with pairwise distance at most ε_ℓ. The final factor has entries at least μ. Taking convex combinations of the product rows gives, for $x, y \in \Delta$ and nonempty w,

$$\mathcal{F}_w(x)_i \geq \mu, \qquad \|\mathcal{F}_w(x) - \mathcal{F}_w(y)\|_1 \leq \varepsilon_{|w|}. \tag{14}$$

In particular every past $H = (\ldots, h_{-1}, h_0)$ has a canonical posterior version

$$\widehat{p}(H) = \lim_{\ell \to \infty} \mathcal{F}_{h_{-\ell+1:0}}(\pi). \tag{15}$$

Two approximants sharing their last ℓ symbols differ by at most ε_ℓ, so this limit is uniform and past-measurable. It belongs to $\mathcal{V} \cap \Delta$, has coordinates at least μ, and satisfies $\widehat{p}(Ha) = \mathcal{F}_a(\widehat{p}(H))$ by continuity. The finite-past approximants are the source's conditional hidden rows; martingale convergence identifies this version with p_H almost surely. Use this version in the ideal coordinates below.

3. Use small local coordinates. Fix

$$\delta = \frac{\mu}{8mL_\Lambda}, \qquad b_0 \geq n \text{ with } \varepsilon_{b_0} \leq \frac{\mu}{128mL_\Pi L_\Lambda}, \qquad \varrho = 1/4. \tag{16}$$

For $u \in \mathcal{U}$ set $c_u = \mathcal{F}_u(\pi) \in \mathcal{V}$ and $x_u(z) = c_u + \delta\Lambda z$. At a block boundary following u, a true posterior y has coordinate $z = \Pi(y - c_u)/\delta$ satisfying

$$\|z\|_1 \leq L_\Pi \varepsilon_{b_0}/\delta \leq 1/16. \tag{17}$$

The lift $y' = x_u(z)$ lies in $\mathcal{V} \cap \Delta$ since $\|\delta\Lambda z\|_1 \leq \mu/(128m)$ and $c_u \geq \mu$. Its difference from y lies in I_0, so (13) gives the next coordinate

$$G_{uv}(z) = \Pi\big(\mathcal{F}_v(c_u + \delta\Lambda z) - c_v\big)/\delta, \qquad |u| = |v| = b_0.$$

For its Taylor coefficients write

$$h_\ell = (T_v\mathbf{1})_\ell, \quad R_\ell = e_\ell T_v/h_\ell, \quad q = c_u \cdot h, \quad f = \mathcal{F}_v(c_u),$$

and define row vectors C_j and scalars γ_j by

$$C_j = \frac{\sum_\ell \lambda_j(\ell) h_\ell (R_\ell - f)}{q}, \qquad \gamma_j = \frac{\delta\lambda_j h}{q}.$$

Direct affine division gives

$$\mathcal{F}_v(c_u + \delta\Lambda z) = f + \frac{\delta \sum_j C_j z_j}{1 + \gamma \cdot z}, \qquad G_{uv}(z) = \frac{\Pi(f - c_v)}{\delta} + \frac{\sum_j \Pi(C_j) z_j}{1 + \gamma \cdot z}. \tag{18}$$

Since $q \ge \mu h_\ell$ and $\|R_\ell - f\|_1 \le \varepsilon_{b_0}$,

$$\sum_j |\gamma_j| \le \frac{\delta m L_\Lambda}{\mu} = \frac{1}{8}, \quad \|C_j\|_1 \le \frac{L_\Lambda \varepsilon_{b_0}}{\mu}, \quad \|\Pi(f - c_v)\|_1 \le L_\Pi \varepsilon_{b_0}.$$

The denominator has no zero on the closed complex polydisc $|z_j| \le 4$. If $\|\cdot\|_{\mathrm{A}}$ is the sum of absolute Taylor coefficients, its geometric expansion on the unit polydisc proves

$$q_0 := \sum_{i=1}^{m} \|G_{uv,i}\|_{\mathrm{A}} \le \frac{L_\Pi \varepsilon_{b_0}}{\delta} + \frac{8 m L_\Pi L_\Lambda \varepsilon_{b_0}}{7\mu} = \frac{64 m L_\Pi L_\Lambda \varepsilon_{b_0}}{7\mu} \le \frac{1}{14}. \tag{19}$$

4. Implement the truncated moments. Retain the monomials z^α, $1 \le |\alpha| \le K$, in m variables. Their number is

$$N_K = \binom{K+m}{m} - 1. \tag{20}$$

Their ideal absolute sum is at most $\prod_j (1 - |z_j|)^{-1} - 1 \le (1 - \|z\|_1)^{-1} - 1 \le 1/15$. Expand $G_{uv}(z)^\alpha = \sum_\nu a_{uv,\alpha\nu} z^\nu$. Take constant coefficient $\mathsf{u}_\alpha = a_{uv,\alpha 0}$ and affine matrix $\mathsf{M}_{\alpha\nu} = a_{uv,\alpha\nu}$ for $1 \le |\nu| \le K$. With $q_i = \|G_{uv,i}\|_{\mathrm{A}}$, absolute convolution gives

$$U + S \le \sum_{\alpha \ne 0} \prod_i q_i^{\alpha_i} = \prod_i (1 - q_i)^{-1} - 1 \le \frac{q_0}{1 - q_0} \le \frac{1}{13} < \frac{1}{3}. \tag{21}$$

Here $U = \sum |\mathsf{u}_\alpha|$, $S = \sum |\mathsf{M}_{\alpha\nu}|$; the last step uses $\prod_i (1 - q_i) \ge 1 - \sum_i q_i$. Lemma 2.4 supplies stochastic maps $K_{u,v}$ on $N = 2^\ell$ labels if $N_K \le N - 1$, or CPTP maps $\mathcal{K}_{u,v}$ on dimension $d = 2^\ell$ if $N_K \le d^2 - 1$. Pad unused Walsh or Pauli features by zero. On every input the maps have a uniform-preparation component of weight at least $2/3$ and contract nonconstant coordinates by $S \le 1/3$. Since $|z_j| \le 1/16 < \varrho$, omitted degrees satisfy

$$\|\tau_{uv}(z)\|_1 \le \varrho^{K+1} \sum_{1 \le |\alpha| \le K} \sum_{|\nu| > K} |a_{uv,\alpha\nu}| \le e_K := \tfrac{1}{3} \varrho^{K+1}. \tag{22}$$

5. One update per symbol, and a stationary device. The memory consists of the label (u, w) of (12) and a register for the monomials. Here u is the last completed block and w is the buffer, with $b_0 \le |w| \le 2b_0 - 1$. When a symbol a is read and $|w| < 2b_0 - 1$, the device appends a to w. Otherwise it splits $wa = vw'$ into two halves of length b_0, applies $K_{u,v}$ or $\mathcal{K}_{u,v}$ to the register, and replaces (u, w) by (v, w'). We call this a *block update*. The classical rule is a stochastic matrix. The quantum rule measures the label, applies the map that the label selects, and writes the new label; it is a channel on every input.

These rules have a stationary realization. Choose a position J_0 uniform on $\{0, \dots, b_0 - 1\}$, independent of the source and of all the device's randomness, and set

$$J_t = (J_0 + t) \bmod b_0, \quad s_t = t - b_0 - J_t, \quad u_t = X_{s_t - b_0 + 1 : s_t}, \quad w_t = X_{s_t + 1 : t}. \tag{23}$$

Thus J_t is the position inside the current block, s_t is the end of the last block that has been used, u_t is that block and w_t is the buffer. When $J_t \ne 0$ the buffer grows by one symbol. When $J_t = 0$ a block update occurs and s_t advances by b_0. The position is stored as $|w_t| - b_0$, so the device needs no clock and no record of older symbols.

Fix one alignment of the blocks. Each matrix $K_{u,v}$ has entries at least $2/(3N)$, so

$$K_{u,v} = \tfrac{1}{2} \mathbf{1} \nu_N + \tfrac{1}{2} \widetilde{K}_{u,v}, \qquad \nu_N = (1/N, \dots, 1/N).$$

Apply the first part of Lemma 2.3 at the times of block updates, and leave the register unchanged between them. The resulting register state depends only on the blocks already used and on independent randomness. Together with the label it is the whole classical memory.

In the quantum case, for each position j at time t, compose the channels of the last k block updates. Lemma 2.3, with contraction $1/3$, gives a limit $\sigma_{t,j}$: starting still earlier changes the present state by at most $2 \cdot 3^{-k}$. The limit is a measurable function of the blocks already used and obeys the update rule. Let $(u_{t,j}, w_{t,j})$ be the labels in (23) when $J_t = j$. The state of the whole memory given the past is

$$\rho_{X_{\le t}} = \frac{1}{b_0} \sum_{j=0}^{b_0-1} |u_{t,j}, w_{t,j}\rangle\langle u_{t,j}, w_{t,j}| \otimes \sigma_{t,j}. \tag{24}$$

Different positions have buffers of different lengths, hence orthogonal labels. The channel for the next symbol sends each term to the term of the next position, with a block update exactly when one is due. It therefore maps (24) to the corresponding state one step later. Nothing outside this density matrix records the position.

The construction commutes with shifts of the observations, of the device's randomness and of the uniform position, so the classical joint law is stationary. The randomness is fresh at each step, which gives the fixed conditional transition probabilities. Given $X_{\le t}$, the remaining randomness is independent of future symbols, which proves (6). In the quantum case a shift permutes the terms of (24) and shifts their limits, so the state is a stationary measurable function of the past. The device keeps only the label and the register.

For the error bound, the exact coordinate at the end s_t of the last used block refers to the posterior given $X_{\le s_t}$, not to the posterior given the later symbols in the buffer as well. Given the whole observed past and the position, the symbols in the buffer do not change the conditional mean of the register, which was last updated at s_t. The exact coordinates follow (13). The stored means follow an affine recursion with contraction $\kappa \le 1/3$ and error $\tau \le e_K$ per block. Lemma 2.5 therefore gives, for each position,

$$E \le e_K/(1 - 1/3) \le \tfrac{1}{2}\varrho^{K+1}. \tag{25}$$

6. Generation on the same memory, and the error. For each label (u, w), expand the posterior at the cut as a function of the stored coordinate:

$$\mathcal{F}_w(c_u + \delta\Lambda z) = f + \sum_{\nu \ne 0} h_{uw,\nu} z^\nu, \qquad f = \mathcal{F}_w(c_u) \ge \mu.$$

Repeating the calculation of Step 3 for $|w| \ge b_0$, without the factor Π/δ, gives

$$\sum_{\nu \ne 0} \|h_{uw,\nu}\|_1 \le \frac{8\delta m L_\Lambda \varepsilon_{b_0}}{7\mu} = t_{b_0} := \varepsilon_{b_0}/7 < \mu/2. \tag{26}$$

The operators $O_{uw,i} = f_i I + \sum_{1 \le |\nu| \le K} h_{uw,i\nu} P_\nu$ are at least $(\mu/2)I$ and sum to I on every input, because each nonconstant coefficient vector sums to zero over i, also after truncation. They form a measurement with n outcomes. Walsh signs give classical readout probabilities with the same properties.

For position J, let y_J be the true posterior at the end of the last used block and $y'_J = x_u(z)$ its representative from Step 3. Put $p'_J = \mathcal{F}_w(y'_J)$. Equation (13) gives $P_{p'_J} = P_{p_H}$ for the present true posterior p_H. The truncated readout has induced ℓ^1 norm $B \le t_{b_0}$ and residual $\epsilon \le t_{b_0}\varrho^{K+1}$. Its mean output $q_{H,J}$ satisfies

$$\|q_{H,J} - p'_J\|_1 \le \tfrac{3}{2} t_{b_0} \varrho^{K+1}, \qquad q_{H,J,i} \ge \mu/2.$$

At the cut the device performs the measurement selected by its label. Since $H_S \geq b_0 \geq n$, choose n orthogonal vectors g_i of the memory, or n classical states; on outcome i the cut channel prepares g_i. With $E_1^{\mathrm{gen}} = |g_1\rangle\langle g_1| + \Pi_\perp$ and $E_i^{\mathrm{gen}} = |g_i\rangle\langle g_i|$ for $i > 1$, the device then runs the source exactly:

$$\mathcal{G}_a(X) = \sum_{i,j=1}^{n} T_a(i,j)\,\mathrm{tr}(E_i^{\mathrm{gen}} X)|g_j\rangle\langle g_j|.$$

These maps are completely positive and their sum is trace preserving on the whole memory. Classically, use the exact transitions on the chosen states and any valid rows elsewhere. Generation uses no memory beyond what was already counted.

After averaging over the device's randomness, $Q_H = b_0^{-1}\sum_J P_{q_{H,J}}$. For each position J apply Lemma 2.5 to the prior p'_J, with $(\tau, \kappa, B, \epsilon, \mu_{\mathrm{read}}) = (e_K, 1/3, t_{b_0}, t_{b_0}\varrho^{K+1}, \mu/2)$. Convexity over the positions then gives the bound for the entire future,

$$\mathcal{R} \leq \frac{9t_{b_0}^2}{2\mu}\,\varrho^{2K+2}.$$

For one fixed alignment of the blocks the device is periodic, not stationary. Mixing uniformly over the b_0 alignments makes it stationary, and the error bound holds for each alignment.

Finally choose the largest power of two $N \leq D/H_S$ or $d \leq D/H_S$; for large D they exceed $D/(2H_S)$. Since $N_K + 1 \leq (K+1)^m$, choose $K + 1 = \lfloor N^{1/m} \rfloor$ classically or $K + 1 = \lfloor d^{2/m} \rfloor$ quantumly. The capacity conditions hold, and these powers give the asserted exponents. For the finitely many remaining $D \geq n$, the constant forecast P_π and an exact n-state generator have risk at most $\max_i \log(1/\pi_i)$. Increasing C_S covers them. Padding unused states or dimensions with valid maps gives devices for every prescribed budget. □

4 Lower bounds from test points

This section proves one lower bound for positive invertible sources whose word-probability rank equals their hidden-state count. Sections 5 and 6 verify this condition and apply the bound.

The argument compares two tables of numbers. Fix an earlier past, some words u_i that may be observed next, some words v_j that may be observed after them, and some words w_ι that may be generated after the cut. The true table lists the probability of w_ι after $u_i v_j$. The device's table lists its own probabilities. The device's table has rank at most r_D, because its forecasts pass through its memory after u_i. If the true table is far from every table of that rank, the device must err.

4.1 Two sets of conditional distributions

Let Δ be the simplex of probability rows and Δ^* the simplex of nonnegative columns with entry sum one. A word $w = a_1 \cdots a_k$ has matrix $T_w = T_{a_1} \cdots T_{a_k}$. It acts on Δ by the posterior update $\mathcal{F}_w$ and on Δ^* by

$$\mathcal{F}_w^*(h) = \frac{T_w h}{\mathbf{1}^{\mathsf{T}} T_w h}.$$

The column $T_w\mathbf{1}$ lists the probability of w from each hidden state, so $\mathcal{F}_w^*(\mathbf{1}/n)$ is this likelihood column scaled to sum one. Define

$$\Omega_{\rightarrow} = \bigcap_{k\geq 1} \bigcup_{|w|=k} \mathcal{F}_w(\Delta), \qquad \Omega_{\leftarrow} = \bigcap_{k\geq 1} \bigcup_{|w|=k} \mathcal{F}_w^*(\Delta^*). \tag{27}$$

We call $\Omega_{\rightarrow}$ the *posterior set*: it consists of the hidden-state distributions that arbitrarily long pasts can produce. We call $\Omega_{\leftarrow}$ the *likelihood set*: it consists of the limiting directions of likelihood columns of long words. We use three facts.

Lemma 4.1. *Let every T_a be strictly positive.*

1. *There is $\theta \in (0,1)$ such that for all nonempty words w, all $x, x' \in \Delta$ and all $h, h' \in \Delta^*$,*

$$\|\mathcal{F}_w(x) - \mathcal{F}_w(x')\|_1 \le 2\theta^{|w|}, \qquad \|\mathcal{F}^*_w(h) - \mathcal{F}^*_w(h')\|_1 \le 2\theta^{|w|}. \tag{28}$$

2. *Both sets are compact, $\mathcal{F}_w(\Omega_{\rightarrow}) \subset \Omega_{\rightarrow}$ and $\mathcal{F}^*_w(\Omega_{\leftarrow}) \subset \Omega_{\leftarrow}$ for every word. The normalized positive left eigenvector of each T_a lies in $\Omega_{\rightarrow}$, and the normalized positive right eigenvector lies in $\Omega_{\leftarrow}$.*

3. *With μ as in (14), $\langle p, h\rangle \ge \mu$ for every $p \in \mathcal{F}_w(\Delta)$ with w nonempty and every $h \in \Delta^*$. Every point of $\Omega_{\rightarrow}$ and of $\Omega_{\leftarrow}$ has strictly positive entries.*

Proof. 1. The first bound is (14) with $\theta \ge 1 - \beta$. Its proof uses only positivity of each factor: the matrices S_k there are stochastic because $h^{(k-1)} = T_{a_k} h^{(k)}$ by definition, and the normalization of $\sum_a T_a$ is not used. The map $\mathcal{F}^*_w$ is the posterior update of the transposed matrices along the reversed word, so the same proof gives the second bound with the constant β^* of the transposes. Take $\theta = \max\{1-\beta, 1-\beta^*\}$.

2. Each union in (27) is a finite union of compact sets, and the unions decrease with k because $\mathcal{F}_{aw}(\Delta) \subset \mathcal{F}_w(\Delta)$. If $p \in \mathcal{F}_{w'}(\Delta)$, then $\mathcal{F}_w(p) \in \mathcal{F}_{w'w}(\Delta)$; letting $|w'|$ grow gives invariance. A normalized left eigenvector ℓ_a satisfies $\ell_a = \mathcal{F}_{a^k}(\ell_a)$ for every k. The statements for $\Omega_{\leftarrow}$ are the same.

3. By (14) the entries of p are at least μ, and the entries of h sum to one. In the same way the entries of $\mathcal{F}^*_a(h)$ are at least $\min_{i,j} T_a(i,j)/(\mathbf{1}^{\mathsf{T}} T_a)_j > 0$. □

4.2 The rank of a device's forecasts

Lemma 4.2 (Forecast tables)**.** *Fix an admitted device, classical or quantum, and a cut. Let $u_1, \dots, u_M$ be distinct words of one length ℓ_1, let $v_1, \dots, v_{M'}$ be distinct words of one length ℓ_2, and let $w_1, \dots, w_k$ be distinct words of one length. For almost every past H_0 ending $\ell_1 + \ell_2$ symbols before the cut, put*

$$X_{i,(j,\iota)} = P_{p_H}(w_\iota), \qquad Y_{i,(j,\iota)} = Q_H(w_\iota), \qquad H = H_0 u_i v_j.$$

These are the true and the generated probabilities that the future begins with w_ι. Then

$$\operatorname{rank} Y \le r_D, \qquad \mathbb{E}\big[D_{\mathrm{KL}}(P_{p_H} \| Q_H) \mid H_0\big] \ge \frac{\varpi_0}{2} \|X - Y\|_F^2, \tag{29}$$

where $\varpi_0 = \min_{i,j} \Pr(u_i v_j \mid H_0)$ and the expectation is over the $\ell_1 + \ell_2$ symbols between H_0 and the cut.

Proof. By Lemma 2.1, for almost every H_0 and all words, the conditional mean memory at H is $\xi_{H_0} K_{u_i} K_{v_j}$, or $\Phi_{v_j}(\Phi_{u_i}(\rho(H_0)))$. The generated law is linear in this mean memory, through the fixed cut channel and the generator. Hence there are columns $c_w \in \mathbb{R}^D$, or Hermitian matrices E_w, depending only on the generated word, with

$$Y_{i,(j,\iota)} = \big\langle \xi_{H_0} K_{u_i},\, K_{v_j} c_{w_\iota} \big\rangle \quad \text{or} \quad Y_{i,(j,\iota)} = \operatorname{tr}\big[\Phi_{u_i}(\rho(H_0))\, \Phi^*_{v_j}(E_{w_\iota})\big].$$

The first factor depends only on the row and lies in a real space of dimension D, respectively D^2. This proves the rank bound.

For each pair (i,j), restricting both laws to the first symbols, as many as the words w_ι have, cannot increase the divergence. Pinsker's inequality in nats then gives

$$D_{\mathrm{KL}}(P_{p_H}\|Q_H) \geq \frac{1}{2}\Big(\sum_w |P_{p_H}(w) - Q_H(w)|\Big)^2 \geq \frac{1}{2}\sum_{\iota=1}^{k}\big(X_{i,(j,\iota)} - Y_{i,(j,\iota)}\big)^2,$$

where the first sum is over all words of the common length of the w_ι. The events $u_i v_j$ are disjoint. Summing over them with their conditional probabilities gives the second bound. □

No assumption on the device enters beyond the model of Section 2. The past H_0 is fixed before the bound is applied, and only the resulting number is averaged over H_0.

4.3 Recovering reciprocal word probabilities

The next identity turns forecasts into a kernel that depends on the source only through $\Omega_\rightarrow$ and $\Omega_\leftarrow$.

Lemma 4.3 (Reciprocal word probabilities)**.** *Let $w_1,\dots,w_n$ be words of one common length whose columns $g_\iota = T_{w_\iota}\mathbf{1}$ form an invertible matrix $G = [g_1\ \cdots\ g_n]$. Let v be a word with T_v invertible and put $\alpha(v) = G^{-1}T_v^{-1}\mathbf{1}$. Then for every $p \in \Delta$,*

$$\sum_{\iota=1}^{n}\alpha_\iota(v)\,\frac{pT_vT_{w_\iota}\mathbf{1}}{pT_v\mathbf{1}} = \frac{1}{pT_v\mathbf{1}}, \qquad \|\alpha(v)\|_2 \leq \sqrt{n}\,\|G^{-1}\|_2\,\big(\max_a \|T_a^{-1}\|_2\big)^{|v|}. \tag{30}$$

Proof. $\sum_\iota \alpha_\iota(v)T_v g_\iota = T_vG\alpha(v) = \mathbf{1}$ and $p\mathbf{1} = 1$. The norm bound uses $\|\mathbf{1}\|_2 = \sqrt{n}$. □

The fractions in (30) are the probabilities of the next words w_ι after a posterior p and an observed word v. A fixed combination of these forecasts therefore equals the reciprocal of the probability of v.

4.4 The lower bound

For finite sets $\mathcal{P} \subset \Omega_\rightarrow$ and $\mathcal{B} \subset \Omega_\leftarrow$ with the same number M of points, define the $M \times M$ matrix

$$\mathsf{R}(\mathcal{P},\mathcal{B}) = \Big[\frac{1}{\langle p,h\rangle}\Big]_{p\in\mathcal{P},\,h\in\mathcal{B}}. \tag{31}$$

Here $\langle p,h\rangle$ is the probability assigned by the hidden distribution p to the normalized likelihood h.

Theorem 4.4 (Lower bound from test points)**.** *Let S have $n \geq 2$ hidden states, strictly positive invertible symbol matrices and word-probability rank $\rho_S = n$. There are constants $c_S > 0$ and $\kappa_S \geq 2$ with the following property. Suppose M-point sets $\mathcal{P} \subset \Omega_\rightarrow$ and $\mathcal{B} \subset \Omega_\leftarrow$ satisfy*

$$\sigma_{\min}\big(\mathsf{R}(\mathcal{P},\mathcal{B})\big) \geq s > 0.$$

Then every classical or quantum device with $r_D < M$ has risk at least $c_S\,(s/M)^{\kappa_S}$.

Proof. Constants. Let θ and μ be as in Lemma 4.1, let $\varpi = \min_{a,i}(T_a\mathbf{1})_i$ and $\Lambda_{\mathrm{inv}} = \max_a \|T_a^{-1}\|_2$. Here $\Lambda_{\mathrm{inv}} \geq 1$, because the least singular value of T_a is at most its spectral radius, which is at most its largest row sum. Let $E_k = \mathrm{span}\{T_w\mathbf{1} : |w| = k\}$. Since $\sum_a T_a\mathbf{1} = \mathbf{1}$, we have $T_w\mathbf{1} = \sum_a T_{wa}\mathbf{1}$, so $E_k \subset E_{k+1}$. Also $E_{k+1} = \sum_a T_aE_k$, so an equality $E_k = E_{k+1}$ persists for all larger indices. As $\dim E_0 = 1$, the chain is constant from index $n-1$ on. The word-probability matrix $[\pi T_uT_v\mathbf{1}]_{u,v}$ has

rank n, so the columns $T_v\mathbf{1}$ span $\mathbb{R}^n$ and $E_{n-1} = \mathbb{R}^n$. Fix words $w_1, \dots, w_n$ of length $n-1$ whose columns $g_\iota = T_{w_\iota}\mathbf{1}$ form an invertible matrix G.

Test words. Write $\mathcal{P} = \{p_i\}$ and $\mathcal{B} = \{h_j\}$, and let ℓ be the least positive integer with

$$\theta^\ell \le \frac{s\mu^2}{8M}. \tag{32}$$

The threshold is below one, since $s \le \|\mathsf{R}\|_2 \le M/\mu$. Choose $u_i \in \mathcal{A}^\ell$ with $p_i \in \mathcal{F}_{u_i}(\Delta)$ and $v_j \in \mathcal{A}^\ell$ with $h_j \in \mathcal{F}^*_{v_j}(\Delta^*)$. For any $p_0 \in \Delta$ put

$$p_i' = \mathcal{F}_{u_i}(p_0), \qquad h_j' = \mathcal{F}^*_{v_j}(\mathbf{1}/n) = \frac{T_{v_j}\mathbf{1}}{\mathbf{1}^{\mathsf{T}} T_{v_j}\mathbf{1}}.$$

By (28), $\|p_i' - p_i\|_1 \le 2\theta^\ell$ and $\|h_j' - h_j\|_1 \le 2\theta^\ell$, so $|\langle p_i', h_j'\rangle - \langle p_i, h_j\rangle| \le 4\theta^\ell$. Both pairings are at least μ by Lemma 4.1. The matrix $\mathsf{R}' = [1/\langle p_i', h_j'\rangle]$ therefore satisfies

$$\|\mathsf{R}' - \mathsf{R}(\mathcal{P}, \mathcal{B})\|_2 \le \frac{4M\theta^\ell}{\mu^2} \le \frac{s}{2}, \qquad \sigma_{\min}(\mathsf{R}') \ge \frac{s}{2}.$$

The words u_i are distinct. Indeed, if $u_i = u_{i'}$ with $i \ne i'$, then $p_i' = p_{i'}'$ and R' has identical rows, contradicting its positive least singular value. Repeated v_j would similarly give identical columns.

The target table. Put $Z_{ij} = 1/(p_i' T_{v_j}\mathbf{1})$. Since $\mathbf{1}^{\mathsf{T}} T_{v_j}\mathbf{1} \le n$, each column of Z is the corresponding column of R' multiplied by a factor at least $1/n$. Hence, for every p_0,

$$\sigma_{\min}(Z) \ge \frac{s}{2n}. \tag{33}$$

Any device. Fix a cut and a past H_0 ending 2ℓ symbols before it, with posterior $p_0 = p_{H_0}$. Apply Lemma 4.2 to the words u_i, v_j and w_ι. By (3),

$$X_{i,(j,\iota)} = \frac{p_i' T_{v_j} T_{w_\iota}\mathbf{1}}{p_i' T_{v_j}\mathbf{1}}, \qquad \Pr(u_i v_j \mid H_0) = p_0 T_{u_i} T_{v_j}\mathbf{1} \ge \varpi^{2\ell}.$$

Let A be the $Mn \times M$ matrix with entries $\mathsf{A}_{(j,\iota),j} = \alpha_\iota(v_j)$ and zeros elsewhere. It depends only on the column words. Lemma 4.3 gives $X\mathsf{A} = Z$ and $\|\mathsf{A}\|_2 = \max_j \|\alpha(v_j)\|_2 \le \sqrt{n}\,\|G^{-1}\|_2 \Lambda_{\mathrm{inv}}^\ell$. Since $\operatorname{rank}(Y\mathsf{A}) \le \operatorname{rank} Y \le r_D < M$ and Z is $M \times M$,

$$\|X - Y\|_F \ge \frac{\|Z - Y\mathsf{A}\|_F}{\|\mathsf{A}\|_2} \ge \frac{\sigma_{\min}(Z)}{\|\mathsf{A}\|_2}.$$

With (29) and (33),

$$\mathbb{E}\big[D_{\mathrm{KL}}(P_{p_H} \| Q_H) \mid H_0\big] \ge \frac{s^2}{8n^3 \|G^{-1}\|_2^2} \Big(\frac{\varpi}{\Lambda_{\mathrm{inv}}}\Big)^{2\ell}.$$

The right-hand side does not depend on H_0. Averaging over H_0 and using stationarity gives the same bound for the risk. Every entry of $\mathsf{R}(\mathcal{P}, \mathcal{B})$ is at most $1/\mu$, so $s \le \sqrt{M}/\mu$ and $s\mu^2/(8M) < 1$. By the minimality of ℓ in (32), also when $\ell = 1$, $\theta^{\ell-1} > s\mu^2/(8M)$, so

$$\Big(\frac{\varpi}{\Lambda_{\mathrm{inv}}}\Big)^{2\ell} \ge \Big(\frac{\varpi}{\Lambda_{\mathrm{inv}}}\Big)^{2} \Big(\frac{s\mu^2}{8M}\Big)^{\kappa_0}, \qquad \kappa_0 = \frac{2\log(\Lambda_{\mathrm{inv}}/\varpi)}{\log(1/\theta)}.$$

The theorem follows with $\kappa_S = \kappa_0 + 2$ and a constant c_S depending only on the source. □

The theorem reduces a lower bound to a question about two sets and one kernel: how many points of $\Omega_\rightarrow$ and $\Omega_\leftarrow$ can be chosen with the reciprocal matrix well conditioned? An exponentially small s costs only a constant factor in the exponent of the final bound.

5 Sources whose posteriors fill an open region

A source has *thick posterior sets* if $\Omega_{\rightarrow}$ has nonempty interior relative to Δ and $\Omega_{\leftarrow}$ has nonempty interior relative to Δ^*.

Theorem 5.1. *Let S have $n \geq 2$ hidden states, strictly positive invertible symbol matrices, and thick posterior sets. Then $\rho_S = n$, and there is $C_S > 0$ such that, for every $D \geq 1$ and $\diamond \in \{\mathrm{cl}, \mathrm{q}\}$,*

$$G_D^{\diamond}(S) \geq e^{-C_S r_D^{1/(n-1)}}.$$

With Theorem 3.1, the minimum classical state count is $\Theta(L_\varepsilon^{n-1})$ and the minimum quantum dimension is $\Theta(L_\varepsilon^{(n-1)/2})$ as $\varepsilon \downarrow 0$, where $L_\varepsilon = \log(1/\varepsilon)$.

The proof places a product grid inside both sets, in coordinates where the reciprocal matrix becomes an explicit kernel.

Lemma 5.2 (A grid kernel)**.** *Let $m \geq 1$, $0 < \varrho \leq (2m)^{-1/2}$ and $q \geq 2$. Put $t_s = \varrho s/q$ for $1 \leq s \leq q$, let $\Gamma_q = \{t_1, \dots, t_q\}^m$ with $M = q^m$ points γ_i, and*

$$\mathsf{H}_{ij} = \frac{1}{1 - \gamma_i \cdot \gamma_j}, \qquad \gamma \cdot \gamma' = \sum_{l=1}^{m} \gamma_l \gamma'_l.$$

Then H is symmetric positive definite and

$$\lambda_{\min}(\mathsf{H}) \geq q^{-2m} \Big(\frac{\varrho}{8e}\Big)^{2m(q-1)}. \tag{34}$$

Proof. Since $0 \leq \gamma_i \cdot \gamma_j \leq m\varrho^2 \leq 1/2$, the geometric and multinomial expansions converge absolutely:

$$\mathsf{H}_{ij} = \sum_{\nu \in \mathbb{N}_0^m} \frac{|\nu|!}{\prod_l \nu_l!} \gamma_i^\nu \gamma_j^\nu.$$

Every weight is at least one, and every term is a positive semidefinite matrix of rank one. Keeping $0 \leq \nu_l < q$ gives $\mathsf{H} \succeq V_m V_m^{\mathsf{T}}$, where $V_m = V^{\otimes m}$ and $V_{sj} = t_s^{j-1}$ for $1 \leq s, j \leq q$.

The columns of V^{-1} are the coefficient vectors of the Lagrange polynomials of the nodes. For fixed s,

$$\prod_{l \neq s} |t_s - t_l| = \Big(\frac{\varrho}{q}\Big)^{q-1} (s-1)!\,(q-s)! \geq \Big(\frac{\varrho}{q}\Big)^{q-1} \frac{(q-1)!}{2^{q-1}} \geq \Big(\frac{\varrho}{4e}\Big)^{q-1},$$

by $\binom{q-1}{s-1} \leq 2^{q-1}$, $(q-1)! \geq ((q-1)/e)^{q-1}$ and $(q-1)/q \geq 1/2$. The coefficients of $\prod_{l \neq s}(x - t_l)$ have absolute sum at most $\prod_l (1 + t_l) \leq 2^{q-1}$. Each column of V^{-1} thus has ℓ^1 norm at most $(8e/\varrho)^{q-1}$, and $\|V^{-1}\|_2 \leq q(8e/\varrho)^{q-1}$. Hence $\sigma_{\min}(V_m) = \sigma_{\min}(V)^m \geq q^{-m}(\varrho/8e)^{m(q-1)}$, and $\lambda_{\min}(\mathsf{H}) \geq \sigma_{\min}(V_m)^2$. □

Proof of Theorem 5.1. Put $m = n - 1$.

Rank. Every point of $\Omega_{\leftarrow}$ is a limit of normalized columns $T_v \mathbf{1}$, by (28). The span of these columns is a subspace containing a relatively open subset of Δ^*, so it is $\mathbb{R}^n$. In the same way the rows πT_u span $\mathbb{R}^n$. The word-probability matrix is the product of these two families, so $\rho_S = n$.

Coordinates. Fix p^* in the relative interior of $\Omega_{\rightarrow}$ and a column h^* whose normalization lies in the relative interior of $\Omega_{\leftarrow}$, scaled so that $\langle p^*, h^* \rangle = 1$. Both are strictly positive. Put $h^*_{\max} = \max_i h^*_i$. For $p \in \Delta$ and nonzero nonnegative columns h define

$$\varphi(p) = p^* - \frac{p}{\langle p, h^* \rangle}, \qquad \psi(h) = \frac{h}{\langle p^*, h \rangle} - h^*.$$

Then $\langle \varphi(p), h^* \rangle = 0$ and $\langle p^*, \psi(h) \rangle = 0$. Expanding $\langle p^* - \varphi(p), h^* + \psi(h) \rangle$ gives

$$\langle p, h \rangle = \langle p, h^* \rangle \, \langle p^*, h \rangle \, \big(1 - \langle \varphi(p), \psi(h) \rangle\big). \tag{35}$$

The pairing between the rows annihilating h^* and the columns annihilated by p^* is nondegenerate: such a row orthogonal to all such columns is a multiple of p^*, and $\langle p^*, h^* \rangle \neq 0$. Choose dual bases of these two m-dimensional spaces, so that $\langle y, \zeta \rangle = \sum_{l=1}^{m} y_l \zeta_l$ in coordinates. The map $p \mapsto p / \langle p, h^* \rangle$ is a homeomorphism of Δ onto the simplex $\{x \geq 0 : \langle x, h^* \rangle = 1\}$. Hence φ maps a neighbourhood of p^* in Δ onto a neighbourhood of zero, and likewise ψ on Δ^*. Choose $\varrho \in (0, (2m)^{-1/2}]$ such that every y with all $|y_l| \leq \varrho$ lies in $\varphi(\Omega_{\rightarrow})$ and every ζ with all $|\zeta_l| \leq \varrho$ lies in $\psi(\Omega_{\leftarrow})$.

Test points. Let $r = r_D$ and let $q \geq 2$ be the least integer with $q^m > r$, so that $q \leq 3r^{1/m}$. For each node γ_i of the grid Γ_q let $p_i \in \Omega_{\rightarrow}$ have coordinates $\varphi(p_i) = \gamma_i$, and let $h_j \in \Omega_{\leftarrow}$ have coordinates $\psi(h_j) = \gamma_j$. By (35),

$$\frac{1}{\langle p_i, h_j \rangle} = a_i b_j \mathsf{H}_{ij}, \qquad a_i = \frac{1}{\langle p_i, h^* \rangle} \geq \frac{1}{h^*_{\max}}, \qquad b_j = \frac{1}{\langle p^*, h_j \rangle} \geq 1.$$

Hence the reciprocal matrix of these $M = q^m$ points has least singular value at least $s = \lambda_{\min}(\mathsf{H}) / h^*_{\max}$. By (34), $s/M \geq e^{-c_1 q}$ for a constant c_1 of the source. Theorem 4.4 gives risk at least $c_S e^{-\kappa_S c_1 q} \geq e^{-C_S r^{1/m}}$ for every device. The orders of the minimum sizes follow from this bound and Theorem 3.1. □

A covering condition implies the hypothesis and survives small perturbations.

Proposition 5.3. *Let $V \subset \Delta$ be compact with nonempty relative interior.*

1. *If $V \subset \bigcup_a \mathcal{F}_a(V)$, then $V \subset \Omega_{\rightarrow}$. Likewise $V \subset \bigcup_a \mathcal{F}^*_a(V)$ for $V \subset \Delta^*$ gives $V \subset \Omega_{\leftarrow}$.*
2. *Let every T_a be invertible, let V lie in the relative interior of Δ, and let $V \subset \bigcup_a \mathcal{F}_a(\operatorname{int} V)$. Then this inclusion holds for all tuples of matrices sufficiently close to $(T_a)_a$. The same is true for the maps $\mathcal{F}^*_a$.*

Proof. 1. Let $y \in V$. Choose a_1 and $y_1 \in V$ with $y = \mathcal{F}_{a_1}(y_1)$, then a_2 and $y_2 \in V$ with $y_1 = \mathcal{F}_{a_2}(y_2)$, and so on. For each k, $y = \mathcal{F}_{a_k \cdots a_1}(y_k)$ lies in $\mathcal{F}_w(\Delta)$ for a word of length k. Hence $y \in \Omega_{\rightarrow}$. For the maps $\mathcal{F}^*_a$ the same choice gives $y = \mathcal{F}^*_{a_1 \cdots a_k}(y_k)$.

2. For invertible T_a, the map $\mathcal{F}_a$ is the restriction of a projective transformation, with inverse $y \mapsto y T_a^{-1} / (y T_a^{-1} \mathbf{1})$ on its image. The sets $O_a = \mathcal{F}_a(\operatorname{int} V)$ are relatively open and cover the compact set V. Choose compact $V_a \subset O_a$ with $V = \bigcup_a V_a$. Each $\mathcal{F}_a^{-1}(V_a)$ is a compact subset of $\operatorname{int} V$. The inverse map depends continuously on the matrix, uniformly on V_a. For matrices close enough to T_a the perturbed inverse is therefore defined on V_a and maps it into $\operatorname{int} V$. Then V_a lies in the perturbed image of $\operatorname{int} V$, and the covering persists. The argument for $\mathcal{F}^*_a$ is identical. □

Proposition 5.4. *Fix $n \geq 2$ and $c, \epsilon > 0$ with $\epsilon(1 + c)(n + c) \leq 1/(2n)$. For $1 \leq a \leq n$ let $s_a = \mathbf{1}^{\mathsf{T}} + c e_a$ and*

$$T_a = \omega\big(I + \epsilon \, s_a^{\mathsf{T}} s_a\big), \qquad \omega = \frac{1}{n + \epsilon(n + c)^2}.$$

These n symmetric matrices are strictly positive and invertible, and their sum is doubly stochastic. The resulting source with n hidden states and n symbols satisfies the strict covering condition of Proposition 5.3(2) in both directions.

Proof. The diagonal entries of T_a are $\omega(1+\epsilon s_a(i)^2)$ and the others are $\omega\epsilon s_a(i)s_a(j) \geq \omega\epsilon$. Also $\det T_a = \omega^n(1+\epsilon\|s_a\|_2^2) > 0$. Since $\sum_a s_a = (n+c)\mathbf{1}^\mathsf{T}$ and $s_a\mathbf{1} = n+c$,

$$\sum_a T_a\mathbf{1} = \omega\big(n+\epsilon(n+c)^2\big)\mathbf{1} = \mathbf{1},$$

and the sum is symmetric.

For $p \in \Delta$ we have $pT_a = \omega(p + \epsilon\,(ps_a^\mathsf{T})\,s_a)$. With $\hat{s}_a = s_a/(n+c) \in \Delta$ and $x_a(p) = \epsilon(n+c)\,ps_a^\mathsf{T}$,

$$\mathcal{F}_a(p) = (1-\tau_a(p))\,p + \tau_a(p)\,\hat{s}_a, \qquad \tau_a(p) = \frac{x_a(p)}{1+x_a(p)} \in \Big(0, \frac{1}{2n}\Big],$$

because $ps_a^\mathsf{T} \leq 1+c$. Thus $\mathcal{F}_a$ moves each point toward $\hat{s}_a$ along a segment.

The points $\hat{s}_1, \ldots, \hat{s}_n$ are affinely independent. Write points of their convex hull as $\sum_a \lambda_a \hat{s}_a$ and let

$$V = \Big\{\sum_a \lambda_a\hat{s}_a :\ \lambda_a \geq \frac{1}{4n},\ \sum_a \lambda_a = 1\Big\}.$$

This set is compact, has nonempty relative interior and lies in the interior of Δ. Let $y \in V$ and choose a with $\lambda_a \geq 1/n$. For $t \geq 1$ put $y(t) = \hat{s}_a + t(y - \hat{s}_a)$. Its coefficients are $t\lambda_b$ for $b \neq a$ and $1 - t(1-\lambda_a)$. Let $t_{\max} = (1 - \frac{1}{2n})^{-1}$. For $1 < t \leq t_{\max}$,

$$t\lambda_b > \lambda_b \geq \frac{1}{4n}, \qquad 1 - t(1-\lambda_a) \geq 1 - \frac{1-1/n}{1-1/(2n)} = \frac{1}{2n-1} > \frac{1}{4n},$$

so $y(t) \in \operatorname{int} V$. The function $\chi(t) = (1-\tau_a(y(t)))\,t$ is continuous on $[1, t_{\max}]$, with $\chi(1) < 1$ and $\chi(t_{\max}) \geq (1-\frac{1}{2n})t_{\max} = 1$. Choose $t^* \in (1, t_{\max}]$ with $\chi(t^*) = 1$. Then $\mathcal{F}_a(y(t^*)) = \hat{s}_a + \chi(t^*)(y - \hat{s}_a) = y$. Hence $V \subset \bigcup_a \mathcal{F}_a(\operatorname{int} V)$.

Since T_a is symmetric, $\mathcal{F}_a^*(h) = \mathcal{F}_a(h^\mathsf{T})^\mathsf{T}$. The transposed set therefore satisfies the same strict covering for the maps $\mathcal{F}_a^*$. □

Corollary 5.5. *For every $n \geq 2$ there is a nonempty open set of n-state sources with n symbols for which the error orders of Theorem 3.1 are attained, with matching lower bounds for all admitted devices.*

Proof. The two propositions give a nonempty open set of sources with positive invertible matrices and thick posterior sets; positivity and invertibility are open conditions. Theorem 5.1 applies to each of them. □

6 Two hidden states

For two hidden states the simplices are intervals, and Theorem 4.4 applies to every invertible source that is not iid. No condition on the shape of the posterior sets is needed.

Theorem 6.1. *Let all symbol matrices of a two-hidden-state source S be positive and invertible. If its stationary observed process is iid, one classical state or one quantum dimension generates its conditional future exactly. Otherwise there are source-dependent $A_S, C_S, c_S > 0$ such that, for every $D \geq 2$,*

$$e^{-C_S r_D} \leq G_D^\diamond(S) \leq A_S e^{-c_S r_D}, \qquad \diamond \in \{\mathrm{cl}, \mathrm{q}\}. \tag{36}$$

Consequently, as $\varepsilon \downarrow 0$, its minimum classical state count is $\Theta(\log(1/\varepsilon))$ and its minimum quantum dimension is $\Theta(\sqrt{\log(1/\varepsilon)})$.

The iid alternative occurs exactly when all symbol matrices share a positive left eigenvector or all share a positive right eigenvector.

We use contrast coordinates: a row $p \in \Delta$ has contrast $x = p_1 - p_2$ and a column $h \in \Delta^*$ has contrast $z = h_1 - h_2$. Then

$$\langle p, h\rangle = \frac{1+xz}{2}, \qquad \|p - p'\|_1 = |x - x'|, \qquad \|h - h'\|_1 = |z - z'|. \tag{37}$$

Lemma 6.2 (Test points on an interval)**.** *Let S be a two-state source with positive invertible symbol matrices. Suppose two symbols have different positive left eigenvectors, and two symbols have different positive right eigenvectors. There is $C > 0$ such that for every $k \geq 1$, with $q = 2^k$, there are q-point sets $\mathcal{P} \subset \Omega_\rightarrow$ and $\mathcal{B} \subset \Omega_\leftarrow$ with $\sigma_{\min}(\mathsf{R}(\mathcal{P}, \mathcal{B})) \geq e^{-Cq}$.*

Proof. Slopes. In the coordinate p_1, the map $\mathcal{F}_a$ is a ratio of two affine functions with derivative $\det T_a/(pT_a\mathbf{1})^2$, and $pT_a\mathbf{1} \leq 1$. In the coordinate h_1, the map $\mathcal{F}_a^*$ has derivative $\det T_a/(\mathbf{1}^\mathsf{T} T_a h)^2$, and $\mathbf{1}^\mathsf{T} T_a h \leq 2$. Put $\lambda_0 = \min_a |\det T_a|/4 > 0$. For all words w,

$$\|\mathcal{F}_w(p) - \mathcal{F}_w(p')\|_1 \geq \lambda_0^{|w|}\|p - p'\|_1, \qquad \|\mathcal{F}_w^*(h) - \mathcal{F}_w^*(h')\|_1 \geq \lambda_0^{|w|}\|h - h'\|_1. \tag{38}$$

Rows. Let a, b be symbols whose normalized left eigenvectors $\ell_a \neq \ell_b$ differ. By (28) the sets $\mathcal{F}_{a^N}(\Delta)$ and $\mathcal{F}_{b^N}(\Delta)$ have diameter at most $2\theta^N$ and contain ℓ_a and ℓ_b. Fix N so large that they are disjoint, with distance $g \in (0, 1]$ at least. Let $A = a^N$ and $B = b^N$. For $\omega = (\omega_1, \ldots, \omega_k)$ in $\{A, B\}^k$ put $p_\omega = \mathcal{F}_{\omega_1 \cdots \omega_k}(\ell_a)$. These points lie in $\Omega_\rightarrow$ by Lemma 4.1. If ω and ω' agree in their last $j-1$ blocks and differ in the block before them, then p_ω and $p_{\omega'}$ are images, under the same word of $N(j-1)$ symbols, of one point of $\mathcal{F}_A(\Delta)$ and one point of $\mathcal{F}_B(\Delta)$. By (38), $\|p_\omega - p_{\omega'}\|_1 \geq g\lambda_0^{N(j-1)}$. For fixed ω there are 2^{k-j} such ω', and $\sum_{j=1}^k (j-1)2^{k-j} = q - k - 1 < q$. With contrasts x_ω,

$$\prod_{\omega' \neq \omega} |x_\omega - x_{\omega'}| \geq g^{q-1}\lambda_0^{Nq} \geq \bigl(g\lambda_0^N\bigr)^q. \tag{39}$$

Columns. Let c, d be symbols whose normalized right eigenvectors differ. The same construction, with $\mathcal{F}^*$ and with the first block of a word applied last, gives points $h_\omega \in \Omega_\leftarrow$ whose contrasts z_ω satisfy (39), after enlarging N and decreasing g if necessary.

The matrix. By (37), $\mathsf{R}(\mathcal{P}, \mathcal{B}) = 2\mathsf{K}$ with $\mathsf{K}_{ij} = 1/(1 + x_i z_j)$, where i, j index the two families. Cauchy's determinant [21, Chapter I, Section 4, Example 6] gives

$$\det\Bigl[\frac{1}{1 + x_i z_j}\Bigr] = \pm\frac{\prod_{i<i'}(x_i - x_{i'})\prod_{j<j'}(z_j - z_{j'})}{\prod_{i,j}(1 + x_i z_j)},$$

and the same formula holds after deleting one row and one column. These identities include zero contrasts by continuity; their denominators $1 + x_i z_j$ remain positive. The quotient of the two determinants is an entry of the inverse:

$$|(\mathsf{K}^{-1})_{ji}| = \frac{\prod_{j'}(1 + x_i z_{j'})\ \prod_{i' \neq i}(1 + x_{i'} z_j)}{\prod_{i' \neq i}|x_i - x_{i'}|\ \prod_{j' \neq j}|z_j - z_{j'}|} \leq \frac{2^{2q-1}}{(g\lambda_0^N)^{2q}},$$

since $0 < 1 + xz < 2$. Hence $\|\mathsf{K}^{-1}\|_2 \leq q\, 2^{2q-1}(g\lambda_0^N)^{-2q}$ and $\sigma_{\min}(\mathsf{R}(\mathcal{P}, \mathcal{B})) \geq e^{-Cq}$ with $C = 2 + 2\log\bigl(2/(g\lambda_0^N)\bigr)$. □

Proof of Theorem 6.1. Theorem 3.1 gives the upper bounds. Let $P = \sum_a T_a$ with stationary row π. For each T_a let ℓ_a and r_a be its positive left and right eigenvectors, with $\ell_a\mathbf{1} = 1$, and let ρ_a be the corresponding eigenvalue.

The iid case. If all ℓ_a coincide, positivity of P forces their common value to be π and $\sum_a \rho_a = 1$. Then for every word $a_1 \cdots a_L$, $\pi T_{a_1} \cdots T_{a_L} \mathbf{1} = \prod_j \rho_{a_j}$, so the output is iid. If all r_a coincide, their common direction is $\mathbf{1}$, the unique positive right eigenvector of the stochastic matrix P. Thus $T_a \mathbf{1} = \rho_a \mathbf{1}$, which gives the same product formula. Conversely, suppose the output is iid. The probability of the next symbol a is an affine function $s_a + t_a x_H$ of the posterior contrast x_H. If every t_a vanishes, all T_a have right eigenvector $\mathbf{1}$. Otherwise the constant conditional probability of a symbol with $t_a \neq 0$ forces x_H to be constant almost surely. Its mean is the contrast of π. The update identity then gives $\pi T_a = \Pr(X_0 = a)\pi$ for every a, because every symbol has positive probability. This proves the eigenvector criterion. An iid future is generated exactly with one state.

The lower bound. Assume the output is not iid. Then $\rho_S = 2$, and by the criterion neither all ℓ_a nor all r_a coincide. Let $r = r_D$ and let $k \geq 1$ be least with $q = 2^k > r$, so that $q \leq 2r$. Lemma 6.2 gives q-point sets with $s \geq e^{-Cq}$. Theorem 4.4 gives risk at least $c_S(e^{-Cq}/q)^{\kappa_S} \geq e^{-C_S r}$ for every device, which is the lower bound in (36). The orders of the minimum sizes follow. □

7 Discussion

A device of size D has forecasts factoring through dimension r_D, and the source asks it to follow a nonlinear recursion. Storing monomials of the posterior does this with error e^{-cK} at a cost of about K^m parameters, where $m + 1$ is the rank of the source. Conversely, well conditioned test points in the posterior set and the likelihood set force every device to have many parameters. When the two sets fill open regions, or when the source is positive, invertible, non-iid and has two hidden states, the two bounds have the same order.

In these cases the quantum dimension is of square-root order in the classical state count. The saving is a factor of two in qubits against bits, at a memory size of order $\log\log(1/\varepsilon)$ bits. It comes from the size of the state space; the devices constructed here measure and re-prepare their register at every block update.

The main open question is what determines the rate when the posterior set is thin. The curve source of Appendix A has rank three and a faster upper bound; no matching lower bound is proved for it. Theorem 4.4 reduces such lower bounds to a question about two sets: how many of their points can be chosen with a well conditioned reciprocal matrix. Beyond the two cases treated here we do not know the answer, in particular when the sets are fractal. Sources with singular symbol matrices are outside Theorem 4.4, because the reciprocal identity uses the inverse matrices. Appendix B shows that this is a substantive boundary: a positive singular two-state source has the faster $\exp[-\Theta(r_D \log(r_D + 1))]$ order. Its proof uses the same feature and forecast-rank methods, without the inverse identity.

The constants in all bounds depend on the source and can be very large, and the bounds are asymptotic. The devices use exact real coefficients. We prove nothing about running time, finite precision or total hardware.

AI assistance and responsibility

OpenAI Codex and Anthropic Claude assisted with mathematical derivations, computational checks, literature comparison and drafting. The author is responsible for the statements, proofs, citations and computations. The recorded reviews are checks made within the project, not external peer review or formal verification.

Data and code

No empirical dataset is used. Scripts in exact and floating-point arithmetic check finite instances of the identities and inequalities in Sections 3–6. They support the written proofs and do not replace them.

A A source whose posterior set is a curve

The source below has three hidden states, rank three and invertible symbol matrices, but its posteriors lie on a curve. Its error is bounded above by Ce^{-cr_D}. No matching lower bound is asserted. It shows that the rank alone does not determine the rate, and that Theorem 5.1 needs its hypothesis.

Proposition A.1. *There is a strictly positive two-symbol source with three hidden states, invertible symbol matrices and word-probability rank three such that, for all $D \geq 3$,*

$$G_D^{\mathrm{cl}} \leq Ce^{-cD}, \qquad G_D^{\mathrm{q}} \leq Ce^{-cD^2}$$

for fixed $C, c > 0$. No matching lower rate is asserted for this source.

Proof. For $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ set

$$S(M) = \begin{pmatrix} a^2 & 2ab & b^2 \\ ac & ad+bc & bd \\ c^2 & 2cd & d^2 \end{pmatrix}.$$

Then $v(p) = (p_1^2, 2p_1p_2, p_2^2)$ satisfies $v(p)S(M) = v(pM)$ and $S(MN) = S(M)S(N)$. Take

$$M_+ = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad M_- = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathsf{N} = S(M_+) + S(M_-) = \begin{pmatrix} 5 & 6 & 2 \\ 3 & 6 & 3 \\ 2 & 6 & 5 \end{pmatrix}.$$

Put $\Lambda = (13 + \sqrt{145})/2$, $b_\star = (\sqrt{145} - 1)/12 \in (9/10, 1)$, $h = (1, b_\star, 1)^{\mathsf{T}}$ and $H = \mathrm{diag}(h)$. Since $\mathsf{N}h = \Lambda h$, the positive matrices

$$T_\pm = \Lambda^{-1} H^{-1} S(M_\pm) H$$

sum to a stochastic matrix. They have determinant $\Lambda^{-3} > 0$; write π for the positive stationary hidden law.

For $p_x = ((1+x)/2, (1-x)/2)$ the invariant curve is

$$q(x) = \frac{v(p_x)H}{v(p_x)h} = \frac{((1+x)^2/2,\ b_\star(1-x^2),\ (1-x)^2/2)}{Z(x)}, \qquad Z(x) = 1 + b_\star + (1-b_\star)x^2.$$

Its source updates are $q(x) \mapsto q(f_\pm(x))$, where $f_+(x) = (x+1)/(5+x)$ and $f_-(x) = (x-1)/(5-x)$. The maps take $[-1, 1]$ into $[-1/3, 1/3]$ and contract there. Normalized rows of long M products therefore coalesce uniformly; their pair squares and mixed products approach the same curve point. So do positive mixtures after H multiplication and normalization. Finite-past filters thus approach $q(f_w(0))$; contraction and martingale convergence identify the full-past posterior $q(x_H)$, with $x_{H\pm} = f_\pm(x_H)$. The invariant interval $I = [-1/4, 1/4]$ contains this filter, and $q_i(x) \geq 1/8$ there since $Z \leq 2$ and the smallest numerator is $9/32$.

Constant runs, also followed by $+$, put both fixed points and f_+ of the negative fixed point in the posterior support. Their curve rows span $\mathbb{R}^3$, as do the effects for the empty word, $+$ and $++$:

$$\det[h, S(M_+)h, S(M_+)^2 h] = 3b_\star(4b_\star^2 - 5) \neq 0.$$

The closed span of finite-past rows contains these limits, so the word-probability rank is three.

We now track powers of the curve parameter. Take the features $r_j(y) = y^j$, $1 \le j \le K$, and write $f_\sigma(z)^j = \sum_{l\ge 0} a^\sigma_{jl} z^l$ for $\sigma = \pm$. The expansion

$$f_\sigma(z) = \frac{\sigma}{5} + \sum_{l\ge 1} \frac{4(-\sigma)^{l-1}}{5^{l+1}} z^l$$

has absolute coefficient sum $2/5$, with constant term of size $1/5$. Convolution gives $\sum_l |a^\sigma_{jl}| \le (2/5)^j$. For the affine update with constants a^σ_{j0} and matrix $\mathsf{M}_{jl} = a^\sigma_{jl}$, $1 \le j, l \le K$, the absolute sums U of the constants and S of the matrix satisfy

$$U \le \sum_{j\ge 1} 5^{-j} = \frac{1}{4}, \qquad S \le \sum_{j\ge 1} \big[(2/5)^j - (1/5)^j\big] = \frac{5}{12}, \qquad U + S \le \frac{2}{3}.$$

For $|y| \le 1/4$ the omitted terms satisfy

$$\sum_{j=1}^{K} \Big| \sum_{l>K} a^\sigma_{jl} y^l \Big| \le 4^{-K-1} \sum_{j\ge 1} (2/5)^j = \frac{1}{6}\, 4^{-K},$$

and $\sum_j |y|^j \le 1/3$. For power-of-two N, d, Lemma 2.4 implements these updates for $K = N - 1$ or $K = d^2 - 1$ features. The mass bounds $U + S \le 2/3$, $S \le 5/12$ and residual $\tau \le (1/6)4^{-K}$ give stationary devices by Lemma 2.3 (classical refresh weight $1/3$), and feature error $E \le (2/7)4^{-K}$ by Lemma 2.5. Store a length-four lag word w: apply the oldest buffered symbol's map, discard it and append the new symbol. Then $y = x_{t-4}$ and $x_H = f_w(y)$, using $16N$ or $16d$ dimensions. Dephasing the quantum word labels before the controlled local update makes it CPTP on all inputs; the remote-past construction remains causal.

To obtain a valid cut on all inputs, set $\rho = 5/4$. On $|z| \le \rho$,

$$|f_\pm(z)| \le 3/5, \qquad |f_\pm(z) - f_\pm(0)| \le 4/15.$$

The $3/5$ disc is invariant and $|f'_\pm(z)| \le L_0 := 25/121$ there. On the same disc $|Z(z)| > 9/5$ and $|q'_i(z)| \le 2$: one may use $|Z'| \le 3/25$, numerator derivatives at most $8/5$, and numerator moduli at most $34/25$. Thus for every length-four word,

$$|q_i(f_w(z)) - q_i(f_w(0))| \le \epsilon_4 := \frac{8}{15} L_0^3, \qquad |z| \le \rho.$$

Writing $q_i(f_w(z)) = a_{iw0} + \sum_{j\ge 1} a_{iwj} z^j$, Cauchy's estimate gives $|a_{iwj}| \le \epsilon_4 \rho^{-j}$ and

$$\sum_{j\ge 1} |a_{iwj}| \le 4\epsilon_4 = \frac{32}{15} \left(\frac{25}{121}\right)^3 < \frac{1}{32}.$$

Also $a_{iw0} \ge 1/8$. The effects $E_{iw} = a_{iw0} I + \sum_{j=1}^{K} a_{iwj} P_j$ are at least $3I/32$ and sum to I, since $\sum_i q_i = 1$ makes all nonconstant coefficient sums vanish. Walsh probabilities have the same margin; direct sums over word labels define the fixed cut.

The readout coefficients satisfy $B \le 3\epsilon_4 \rho^{-1}$ and the truncation residual is at most $3\epsilon_4 \sum_{j>K} 5^{-j}$ in ℓ^1. Together with the feature error above these give $|\widehat{q}_{H,i} - q_i(x_H)| \le 2\, 4^{-K}$ and $\widehat{q}_{H,i} \ge 3/32$.

The fixed cut prepares three existing labels or orthogonal vectors and runs the exact source, as in Theorem 3.1, giving $Q_H = P_{\widehat{q}_H}$. Lemma 2.2 gives the sharper coordinatewise form of (11):

$$D_{\mathrm{KL}}(P_{q(x_H)} \| P_{\widehat{q}_H}) \le \sum_{i=1}^{3} \frac{(q_i(x_H) - \widehat{q}_{H,i})^2}{\widehat{q}_{H,i}} \le 128\, 16^{-K}.$$

For $D \ge 64$, take the largest power of two N or d at most $D/16$, so both exceed $D/32$ and are at least four. The resulting sufficient bounds are

$$G_D^{\mathrm{cl}} \le 2048 e^{-(\log 16) D/32}, \qquad G_D^{\mathrm{q}} \le 2048 e^{-(\log 16) D^2/1024}.$$

Padding as in Section 2.6 fits every D. For $3 \le D < 64$, a device that ignores its input, starts from π and runs the exact three-state source has risk at most $\max_i \log(1/\pi_i)$. Enlarging the prefactor proves the claimed bounds for all $D \ge 3$. □

B A singular source with a faster optimal rate

Invertibility is a substantive restriction in the two-state classification. The following positive source has two singular symbol matrices and a different optimal rate. The construction uses the same feature updates and the converse uses the same forecast-rank principle as the main results.

Proposition B.1. *In hidden-state order $(+,-)$, let S_{fix} have symbol matrices*

$$\begin{aligned} T_{A_+} &= \frac{1}{16}\begin{pmatrix} 3 & 2 \\ 1 & 2 \end{pmatrix}, \quad T_{A_-} = \frac{1}{16}\begin{pmatrix} 2 & 1 \\ 2 & 3 \end{pmatrix}, \\ T_{B_+} &= \frac{1}{16}\begin{pmatrix} 3 & 1 \\ 3 & 1 \end{pmatrix}, \quad T_{B_-} = \frac{1}{16}\begin{pmatrix} 1 & 3 \\ 1 & 3 \end{pmatrix}. \end{aligned} \tag{40}$$

There are constants $c, C > 0$ such that, for every $D \ge 2$,

$$e^{-C r_D \log(r_D+1)} \le G_D^{\diamond}(S_{\mathrm{fix}}) \le C e^{-c r_D \log(r_D+1)}, \qquad \diamond \in \{\mathrm{cl}, \mathrm{q}\}. \tag{41}$$

Consequently, with $L_\varepsilon = \log(1/\varepsilon)$,

$$M_{\mathrm{cl}}(\varepsilon; S_{\mathrm{fix}}) = \Theta\!\left(\frac{L_\varepsilon}{\log L_\varepsilon}\right), \qquad M_{\mathrm{q}}(\varepsilon; S_{\mathrm{fix}}) = \Theta\!\left(\sqrt{\frac{L_\varepsilon}{\log L_\varepsilon}}\right).$$

Proof. The matrices are strictly positive and sum to $\frac{1}{16}\left(\begin{smallmatrix} 9 & 7 \\ 7 & 9 \end{smallmatrix}\right)$, whose stationary hidden row is uniform. For posterior contrast x, direct multiplication gives

$$\Pr(B_\tau \mid x) = \tfrac{1}{4}, \quad x \mapsto \tau/2, \qquad \Pr(A_\sigma \mid x) = \frac{4 + \sigma x}{16}, \quad x \mapsto F_\sigma(x) := \frac{x}{4 + \sigma x}, \qquad \sigma, \tau \in \{\pm 1\}. \tag{42}$$

For contrast x, write $\pi_x = ((1+x)/2, (1-x)/2)$ and $P_x = P_{\pi_x}$. Thus $|x_H| \le 1/2$ almost surely. Indeed, the probability of no B symbol in the last k observations is 2^{-k}, so a most recent reset exists almost surely and fixes the posterior at $\pm 1/2$. The two reset contrasts give different next-A_+ probabilities, so the output is not iid and its word-probability rank is two.

Moment updates. For $K \ge 1$ put $h_1 = 1$, $h_j = 1/2$ for $j \ge 2$, and retain $r_j(x) = h_j x^j$, $1 \le j \le K$. Their reference vectors have ℓ^1 norm at most $3/4$ on $[-1/2, 1/2]$. The power series of $r_j(F_\sigma(x))$ has truncated update matrix

$$A_\sigma(j, \ell) = \frac{h_j}{h_\ell}(-\sigma)^{\ell - j}\binom{\ell - 1}{j - 1} 4^{-\ell} \quad (\ell \ge j), \qquad A_\sigma(j, \ell) = 0 \quad (\ell < j). \tag{43}$$

Its absolute column sum is $1/4$ for $\ell = 1$ and $(2^{\ell-1}+1)4^{-\ell}$ for $\ell \geq 2$. Hence $\sum_{j,\ell} |A_\sigma(j,\ell)| < 7/12$. Lemma 2.4 implements these updates as stochastic maps on N states with $K = N-1$, or as globally CPTP maps in dimension d with $K = d^2-1$, for powers of two $N, d \geq 2$. The reference preparations are valid by the same lemma.

An exact initial part and the complete memory cost. Choose $J \in \mathbb{N}_0$. Use one extra label (τ, w) for each $\tau \in \{\pm 1\}$ and continuation word w with $0 \leq |w| < J$. It represents the exact contrast $x_{\tau,w} = F_w(\tau/2)$. There are

$$H_J = 2(2^J - 1)$$

such labels. A reset prepares $(\tau, \varnothing)$, continuations move to its children, and the Jth continuation prepares the moment block exactly at the resulting contrast. For $J = 0$ a reset directly prepares the block at $\tau/2$. Subsequent continuations use (43). The total memory is $H_J + N$ classical states or quantum dimension $H_J + d$. Quantumly this is a direct sum of orthogonal initial labels and the d-dimensional block. Dephase between these components, move or prepare each initial component with its input weight, and apply its CPTP update to the block component. These operations are CPTP on every input, and the current label and observed symbol determine every update.

Since $|F_\sigma(x)| \leq (2/7)|x|$ on $[-1/2, 1/2]$, put $a = (2/7)^J$. Every true contrast while the memory is in the block has magnitude at most $a/2$. The omitted power-series terms satisfy

$$\begin{aligned}\tau_K &\leq \sum_{\ell > K} \sum_{j=1}^{K} h_j \binom{\ell-1}{j-1} (a/8)^\ell \\ &\leq \frac{1}{2} \sum_{\ell > K} (a/4)^\ell \leq \frac{1}{6} a^{K+1} 4^{-K}.\end{aligned}$$

Block entry is exact. Lemma 2.5 with contraction $7/12$ therefore bounds the first-feature error by

$$|c_H - x_H| \leq \frac{2}{5} a^{K+1} 4^{-K}. \tag{44}$$

Here c_H is the actual first mean feature in the block and the exact contrast on an initial label. Moreover $|c_H| \leq 1/2$: preparations have this property, and the first absolute row sum of A_σ is at most $5/12$, whereas every physical feature has absolute value at most one.

These are stationary causal devices. At each integer time start at the most recent observed reset, prepare its prescribed memory, and apply the finitely many subsequent continuation updates. Classically use fresh coins independent of the entire source; quantumly compose the specified channels. The classical state is a function of the observed past and these coins; the quantum density is a function of the observed past alone. Both constructions commute with time shifts and obey the fixed online rules. This constructs the stationary memory rather than supplying the source posterior to the device. The reset age is not an additional stored variable: before block entry it is in the initial label, and after entry the same update is used at every continuation.

Generation on the same register. Choose existing block vectors $g_\pm$ with $P_1 g_\pm = \pm g_\pm$ and define on the full quantum register

$$O = \sum_{\tau,w} x_{\tau,w} |\tau, w\rangle\langle \tau, w| \ \oplus \ P_1.$$

Its norm is at most one. The instrument

$$\mathcal{G}_b(X) = \sum_{i,j \in \{\pm 1\}} T_b(i,j) \operatorname{tr}\left(\frac{I + iO}{2} X \right) |g_j\rangle\langle g_j| \tag{45}$$

is CP and sums to a trace-preserving map. On input with mean $c = \mathrm{tr}(OX)$ it emits the exact source from hidden probabilities $(1 \pm c)/2$, and stores each destination hidden state in g_j. Classically assign each initial label its exact contrast and each block label its first Walsh sign; use the analogous stochastic instrument and two existing opposite-sign block labels for the destinations. Both generators therefore have $Q_H = P_{c_H}$ without adding persistent memory. Lemma 2.2 gives $D_{\mathrm{KL}}(P_x \| P_c) \le (x-c)^2/(1-c^2) \le (4/3)(x-c)^2$ on the reached interval. With (44), this gives

$$\mathcal{R} \le C_0 a^{2K+2} 16^{-K} \tag{46}$$

for an absolute constant C_0.

For $D \ge 16$, let D_0 be the largest power of two at most D and choose $J = \log_2 D_0 - 2$, with $N = D_0/2$ or $d = D_0/2$. Then $H_J = D_0/2 - 2$, so the full memory costs $D_0 - 2 \le D$, the block size exceeds $D/4$, and $J \ge (\log_2 D)/4$. Thus $K + 1 \ge r_D/16$ in both cases. The factor a^{2K+2} in (46) gives the upper bound in (41), since $\log(r_D + 1) = O(\log D)$. Padding fits the remaining dimensions. For the finitely many budgets $2 \le D < 16$, ignoring the observations and running the exact source from its uniform hidden prior uses two states or dimensions and has risk at most $\log 2$. Increasing the prefactor covers these budgets.

A converse at one cut. Put $r = r_D$, $M = r + 2$, $k = \lceil \log_2 M \rceil$ and $\xi = 4^{-k}$. Choose M distinct sign words u_i and M distinct sign words v_j, all of length k. Write $s_w = \sum_{\ell=1}^{k} \sigma_\ell 4^{-\ell}$ for a sign word w. Fix almost any full past H_0 before the observed word $B_+ A_{u_i} A_{v_j}$. The reset and $1/F_\sigma(x) = 4/x + \sigma$ give a true cut contrast

$$X_{ij} = \frac{\xi^2}{2 + s_{u_i} + \xi s_{v_j}}.$$

This does not depend on H_0. The probability of each tested word conditional on H_0 is

$$\Pr(B_+ A_{u_i} A_{v_j} \mid H_0) = \frac{\xi^2}{4}\left(1 + \frac{s_{u_i} + \xi s_{v_j}}{2}\right) \ge \frac{5}{24}\xi^2.$$

Apply Lemma 4.2 with row words $B_+ A_{u_i}$, column words A_{v_j}, and generated word A_+. Its true table is $\mathsf{P} = \frac{1}{4}\mathbf{1}\mathbf{1}^{\mathsf{T}} + X/16$, and the device table Q has rank at most r. Consequently $\widetilde{Y} = 16\mathsf{Q} - 4\mathbf{1}\mathbf{1}^{\mathsf{T}}$ has rank at most $r + 1 = M - 1$. The distance of X from such matrices is at least $\sigma_{\min}(X)$. The same lemma, followed by averaging its scalar bound over H_0, yields

$$\mathcal{R} \ge c_0 \xi^2 \sigma_{\min}(X)^2 \tag{47}$$

for an absolute $c_0 > 0$. In particular this argument does not require invertibility of the reset matrices.

For completeness, write $X = \xi^2 \mathsf{C}$ with $\mathsf{C}_{ij} = 1/(a_i + b_j)$, $a_i = 2 + s_{u_i}$ and $b_j = \xi s_{v_j}$. Distinct sign words have s-values separated by at least $4^{-k} = \xi$: the first differing sign dominates the remaining tail. Hence the a_i have gaps at least ξ and the b_j have gaps at least ξ^2. All sums $a_i + b_j$ lie in $(5/3, 7/3)$. The Cauchy inverse identity, also obtained by taking ratios of minors and the determinant in Lemma 6.2, gives

$$|(\mathsf{C}^{-1})_{ji}| = \frac{\prod_l |a_i + b_l| \prod_{l \ne i} |a_l + b_j|}{\prod_{l \ne i} |a_i - a_l| \prod_{l \ne j} |b_j - b_l|} \le 3^{2M-1} \xi^{-3(M-1)}.$$

It follows that

$$\sigma_{\min}(X) \ge \frac{\xi^{3M-1}}{M 3^{2M-1}} \ge e^{-C_1 M \log M},$$

because $k = O(\log M)$. Substitution into (47), with $M = r + 2$, proves the lower bound in (41). Finally, inverting $r \log(r+1) \asymp L_\varepsilon$ gives the stated memory orders. □

References

[1] J. Thompson, A. J. P. Garner, V. Vedral and M. Gu, Using quantum theory to simplify input–output processes, *npj Quantum Information* **3**, 6 (2017). doi:10.1038/s41534-016-0001-3.

[2] L. Li, D. Qiu, X. Zou, L. Li, L. Wu and P. Mateus, Characterizations of one-way general quantum finite automata, *Theoretical Computer Science* **419**, 73–91 (2012). doi:10.1016/j.tcs.2011.10.021.

[3] A. J. P. Garner, Q. Liu, J. Thompson, V. Vedral and M. Gu, Provably unbounded memory advantage in stochastic simulation using quantum mechanics, *New Journal of Physics* **19**, 103009 (2017). doi:10.1088/1367-2630/aa82df.

[4] J. B. Ruebeck, R. G. James, J. R. Mahoney and J. P. Crutchfield, Prediction and generation of binary Markov processes: Can a finite-state fox catch a Markov mouse?, *Chaos* **28**, 013109 (2018). doi:10.1063/1.5003041.

[5] S. E. Marzen and J. P. Crutchfield, Nearly maximally predictive features and their dimensions, *Physical Review E* **95**, 051301(R) (2017). doi:10.1103/PhysRevE.95.051301.

[6] F. Ghafari, N. Tischler, J. Thompson, M. Gu, L. K. Shalm, V. B. Verma, S. W. Nam, R. B. Patel, H. M. Wiseman and G. J. Pryde, Dimensional Quantum Memory Advantage in the Simulation of Stochastic Processes, *Physical Review X* **9**, 041013 (2019). doi:10.1103/PhysRevX.9.041013.

[7] E. R. Anschuetz, H.-Y. Hu, J.-L. Huang and X. Gao, Interpretable Quantum Advantage in Neural Sequence Learning, *PRX Quantum* **4**, 020338 (2023). doi:10.1103/PRXQuantum.4.020338.

[8] 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, *Physical Review A* **108**, 022411 (2023). doi:10.1103/PhysRevA.108.022411.

[9] R. Sundar and T. J. Elliott, Dimension Reduction for Quantum Adaptive Agents, preprint (2026), arXiv:2607.19156v2.

[10] J. Chen, A. Gupta, C. Yang, A. Neufeld, J. Thompson, T. J. Elliott and M. Gu, Quantum-Enhanced Inference of Conditional Future Probabilities with Reduced Memory Cost, preprint (2026), arXiv:2610.10756v1.

[11] P. Feng and C. Meng, Recursive memory for prediction and generation, preprint (2026), arXiv:2611.13760v1.

[12] Z. Chen and J. Wu, The State Cost of Classical Simulation of One-Way General Quantum Finite Automata, preprint (2026), arXiv:2604.07058v2.

[13] L. Banchi, Accuracy vs Memory Advantage in the Quantum Simulation of Stochastic Processes, *Machine Learning: Science and Technology* **5**, 025036 (2024). doi:10.1088/2632-2153/ad444a.

[14] R. Sundar and T. J. Elliott, Quantum dimension reduction of hidden Markov models, *Physical Review A* **113**, 062461 (2026). doi:10.1103/plt9-cm1x.

[15] T. J. Elliott, Memory compression and thermal efficiency of quantum implementations of nondeterministic hidden Markov models, *Physical Review A* **103**, 052615 (2021). doi:10.1103/PhysRevA.103.052615.

[16] A. Monràs and A. Winter, Quantum learning of classical stochastic processes: The Completely-Positive Realization Problem, *Journal of Mathematical Physics* **57**, 015219 (2016). doi:10.1063/1.4936935.

[17] M. Fanizza, J. Lumbreras and A. Winter, Quantum Theory in Finite Dimension Cannot Explain Every General Process with Finite Memory, *Communications in Mathematical Physics* **405**, 50 (2024). doi:10.1007/s00220-023-04913-4.

[18] M. M. Wolf and D. Pérez-García, Assessing Quantum Dimensionality from Observable Dynamics, *Physical Review Letters* **102**, 190504 (2009). doi:10.1103/PhysRevLett.102.190504.

[19] B. Balle, P. Panangaden and D. Precup, Singular value automata and approximate minimization, *Mathematical Structures in Computer Science* **29**, 1444–1478 (2019). doi:10.1017/S0960129519000094.

[20] A. D. Kara and S. Yüksel, Near Optimality of Finite Memory Feedback Policies in Partially Observed Markov Decision Processes, preprint, version 2 (2022), arXiv:2010.07452v2.

[21] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, 2nd ed., Oxford University Press, 1995.

[22] Z. Gao, Time Reversal and the Memory Cost of Prediction, Hexagon preprint 2611.00969v1 (2026), hexagonmath.org/2611.00969v1.