PREDICTIVE MEMORY COMPLEXITY: PERSISTENT-STATE REQUIREMENTS FOR CONTINUAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual world models accumulate an ever-growing history of experience, while the memory that can be persistently retained remains fundamentally bounded. This mismatch raises a fundamental question:how much information from the past must remain encoded in persistent state to preserve future predictive accuracy? We introduce *predictive memory complexity* (PMC), the minimum persistent-state capacity required to attain a prescribed prediction risk. (i) Under stationary fully observed Gaussian dynamics, Theorem 1 gives the exact finite-memory law $R_B^\star = q + a^2 D_{2^B}(v)$, showing that arbitrary recursive memory reduces to quantization of the current predictive state and yielding $B^\star(d) = \frac{1}{2}\log_2 \frac{a^2 v}{d-q} + O(1)$. (ii) Removing stationarity, Theorem 2 shows that memory tracks the scale of the predictive state: $B_t^\star(d) = \frac{1}{2}\log_2 v_t + O(1)$, producing bounded, logarithmic, and linear growth for stable, critical, and unstable dynamics, respectively. (iii) Under partial observability, Theorem 3 bounds the excess prediction risk between quantization of the Bayesian predictive state and the better of two finite-memory mechanisms—history retention and recursive belief representation—thereby separating inferential from representational memory. (iv) Beyond Gaussianity, Theorem 4 preserves the exact quantization reduction while introducing the entropy-power factor $\beta = e^{-2D_{\mathrm{KL}}(P_X \| \mathcal{N}(\mu,v))}$, showing that equal variance need not imply equal finite-memory difficulty. (v) Theorem 5 specializes the theory to an isotropic KV-cache model: with $R = B/(dL)$ bits per scalar, $D_L^\star \geq \sigma_\eta^2/(1-\rho^2 2^{-2R})$, and half of the asymptotically available predictive gain occurs at $R_{\mathrm{onset}} = \frac{1}{2}\log_2(1/(1-\rho^2))$. Measurements from Transformer key representations provide an empirical scale check for the stylized cache regime. The results show that persistent memory is a property of the predictive process rather than merely of the predictor architecture: its required capacity is jointly governed by predictive-state geometry, dynamical expansion, and inferential uncertainty. This provides a principled basis for understanding memory in continual world models, including what predictive information must persist, how accurately it must be represented, and when fixed-capacity memory is fundamentally insufficient.

1 INTRODUCTION

Sequential prediction requires information from the past to remain available as new observations arrive. This principle underlies recurrent models, state-space models, predictive-state representations, and learned world models (Littman et al., 2001; Gu et al., 2022; Gu & Dao, 2024; Ha & Schmidhuber, 2018; Hafner et al., 2020). Recent advances in long-context modeling, structured state-space models, and continual world modeling have further shifted attention from how past information is accessed to what prediction-relevant information must persist in the model's internal state over time. Classical real-time coding, causal rate–distortion theory, and data-rate results provide sharp answers to a closely related question: how much information must be transmitted over time to sustain a desired reconstruction or prediction accuracy. For Gauss–Markov sources, this viewpoint yields explicit rate–distortion laws and critical rates determined by dynamical expansion.

Persistent memory imposes a different operational constraint. A causal receiver may accumulate information transmitted over many steps, whereas a finite-memory predictor must repeatedly compress all prediction-relevant information into a bounded mutable state. Communication rate, context

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

length, and model size therefore do not directly determine how much data-dependent information must remain stored between successive predictions.

This distinction becomes particularly important beyond the simplest stationary fully observed setting. Three structural changes alter the memory problem itself. First, the predictive state need not have Gaussian or otherwise regular geometry, so finite-state approximation is governed by the distribution of the predictive quantity rather than by state cardinality alone. Second, under nonstationary dynamics the scale of the predictive state can expand over time, creating a growing memory requirement even when the current state is directly observed. Third, under partial observability the predictive state is no longer given: it must first be inferred from history and then represented by a finite persistent state. These regimes turn persistent memory from a static architectural choice into a dynamical and inferential resource.

To formalize this resource, consider a predictor with mutable state S_t satisfying $|\mathcal{S}| \leq 2^B$. All data-dependent influence of the observation history on future predictions must pass through S_t; fixed parameters, codebooks, source parameters, and predetermined update rules are treated as read-only specification. For a target long-run prediction risk d, define the *predictive memory complexity* (PMC) $B^\star(d) := \min\{B : \bar{R}^\star_B \leq d\}$, where $\bar{R}^\star_B$ is the optimal long-run risk achievable with B persistent bits. Under partial observability, where even an unrestricted-history predictor incurs irreducible uncertainty, the corresponding excess-risk complexity is $B^\star_{\mathrm{ex}}(\Delta) := \min\{B : \bar{R}^\star_B - \bar{R}^\star_\infty \leq \Delta\}$. PMC quantifies the minimum persistent mutable state required to achieve a target level of predictive performance.

The five theorems form a single progression rather than five independent results.

1). Theorem 1 establishes the core predictive-memory law, relating a finite information budget, dynamical expansion, and achievable prediction error.

2).Theorems 2–4 then extend this framework along three structural directions: Theorem 2 removes stationarity and characterizes dynamical memory growth; Theorem 3 removes full observability and introduces inferential memory; and Theorem 4 removes Gaussianity and shows how distributional geometry enters finite-memory difficulty.

3). Theorem 5 translates the resulting theory into an operational memory threshold for KV-cache compression.

Thus the paper develops one theoretical framework, three structural extensions—causality, distributional robustness, and high-dimensional dynamics—and one concrete application. These results close the gap between classical rate–distortion characterizations of information flow and a predictive-memory view of continual prediction: they identify when a finite information budget is sufficient, how the required budget changes under causality, non-Gaussian uncertainty, and dynamical instability, and how these limits translate into an explicit memory threshold in a modern sequence-model setting.

2 Related Work

Real-time coding, causal prediction, and data-rate limits. Classical real-time source coding and sequential quantization study how information should be encoded causally for stochastic sources (Witsenhausen, 1979; Borkar et al., 2001). Related work extends causal coding to partially observed Markov sources and connects nonanticipative rate–distortion formulations with filtering (Yüksel, 2013; Charalambous et al., 2014). For linear dynamical systems, causal rate–distortion and data-rate results characterize communication requirements for reconstruction, prediction, or stabilization (Derpich & Østergaard, 2012; **?**; Nair & Evans, 2004; Sahai & Mitter, 2006). These theories constrain information transmitted over time and generally permit previously received information to remain available at the decoder or controller. They therefore do not directly characterize the minimum amount of *retained, data-dependent state* that must persist between successive predictions.

Finite-state prediction and bounded memory. A separate literature studies prediction and decision-making with explicitly bounded memory. Early work analyzed learning with finite memory and the value of randomization (Hellman & Cover, 1970; 1971), while finite-state and universal prediction methods considered sequential prediction under restricted internal representations (Feder et al., 1992). Bounded-memory effects have also been studied in hidden-state and changing en-

vironments (Monte & Said, 2010; 2013; Chatterjee et al., 2022), and memory lower bounds have been derived for continual learning across task sequences (Chen et al., 2022). These formulations establish that limited memory can fundamentally constrain sequential performance, but they do not provide a common risk-indexed measure of the persistent state required to represent, track, or infer a predictive state across different stochastic regimes.

Predictive states and long-memory sequence models. Predictive-state representations characterize dynamical systems through statistics sufficient for predicting future observations (Littman et al., 2001; Singh et al., 2004; James & Singh, 2004). Modern recurrent and state-space architectures similarly maintain compressed internal states for long-range sequence modeling (Gu et al., 2022; Gu & Dao, 2024). Recent work has studied stable prediction, robust long-memory parameterizations, length generalization, and combinations of fading memory with retrieval (Ziemann et al., 2025; Yu et al., 2025; Buitrago & Gu, 2025; Nunez et al., 2025). These lines of work primarily characterize representation structure, parameterization, or architectural mechanisms. There is still a lack of direct characterization of the mutable predictive-state capacity required to achieve a prescribed prediction risk.

Non-Gaussian information measures. Shannon rate–distortion theory relates information and reconstruction error (Shannon, 1948; Berger, 1971; Cover & Thomas, 2006), while Rényi divergence and Sibson α-information provide broader families of information measures (Csiszár, 1995; Sibson, 1969; van Erven & Harremoës, 2014). These tools characterize information geometry under increasingly general distributional assumptions, but information measures alone do not specify the cardinality of a persistent finite-state representation. In particular, second-order quantities such as variance need not determine the difficulty of finite-state approximation once the predictive-state distribution is non-Gaussian.

KV-cache compression and long-context inference. Long-context sequence models create an explicit storage bottleneck through their key–value cache. Existing methods such as H2O and StreamingLLM reduce the effective cache or context requirement through selective retention and structured context management (Zhang et al., 2023; Xiao et al., 2024). Related recent work also studies long-context state-space models (Po et al., 2025). These approaches address practical mechanisms for sustaining long-context inference, but they do not directly express cache capacity as a prediction-risk-indexed persistent-state requirement determined jointly by temporal dependence, representation dimension, and retained sequence length.

3 Predictive Memory Complexity: Problem Formulation

Memory is commonly characterized through architecture-specific quantities such as state dimension, context length, or communication rate, leaving unclear how much persistent state is actually required for a given prediction accuracy. We study continual prediction under this finite persistent-memory constraint: all data-dependent information that survives across time must pass through a state S_t. Given an observation O_t, the predictor updates S_t and uses it to predict a future target V_{t+1}. When the predictive state is directly observed, finite memory limits how accurately it can be represented; under partial observability, memory must additionally preserve the history needed to infer it. These two regimes motivate a common resource question: the minimum persistent mutable capacity required to achieve a prescribed prediction risk. We formalize this quantity as predictive memory complexity (PMC), which characterizes the minimum persistent-state capacity required to achieve a prescribed prediction risk.

3.1 How Much Persistent-State Capacity Is Required to Achieve a Target Prediction Risk?

The key question is to determine the minimum persistent-state capacity B required to sustain a target prediction risk d. A predictor with B bits of persistent memory maintains a state

$$S_t = U_t(S_{t-1}, O_t, \xi_t), \qquad S_t \in \mathcal{S}, \qquad |\mathcal{S}| \le 2^B, \tag{1}$$

and produces the prediction $\widehat{V}_{t+1} = g_t(S_t)$. Here ξ_t denotes fresh randomness independent of the source. Such randomness may affect the current update, but it is not available at later times unless its effect is encoded in S_t.

The budget B measures only *mutable, data-dependent state carried across time*. Fixed model parameters, codebooks, source parameters, and predetermined update rules are treated as read-only specification and are not included in B. No additional observation history, residual, counter, or data-dependent auxiliary state may be retained outside S_t. If U_t or g_t varies with time, this variation is predetermined and does not encode the realized observation history.

Let Π_B denote the class of all predictors satisfying this persistent-state constraint. A predictor may use deterministic or randomized updates, but every influence of the past on future predictions must pass through at most B persistent bits.

3.2 What Determines How Persistent-State Requirements Change?

The second key question is to identify the factors that determine persistent-state requirements and characterize how the minimum persistent-state capacity changes with the dynamics, observability, and state distribution of the predictive process. For a loss function ℓ, define the finite-horizon average prediction risk of a predictor $\pi \in \Pi_B$ by

$$R_T(\pi) = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\ell\Big(V_{t+1}, \widehat{V}_{t+1}\Big)\right]. \tag{2}$$

The best risk achievable with B persistent bits is $R^\star_{B,T} := \inf_{\pi \in \Pi_B} R_T(\pi)$.

For stationary or asymptotic questions, define the long-run risk as $\overline{R}^\star_B := \inf_{\pi \in \Pi_B} \limsup_{T\to\infty} R_T(\pi)$.

For a target risk d, the finite-horizon predictive memory complexity is $B^\star_T(d) := \min\{B \in \mathbb{N}_0 : R^\star_{B,T} \le d\}$. For long-run prediction, we define

$$B^\star(d) = \min\left\{B \in \mathbb{N}_0 : \overline{R}^\star_B \le d\right\}. \tag{3}$$

Under partial observability, it is often more informative to measure performance relative to the unrestricted Bayesian predictor. Let $\overline{R}^\star_\infty$ denote the optimal long-run risk when the entire observation history may be used. We define the *excess-risk predictive memory complexity* as

$$B^\star_{\mathrm{ex}}(\Delta) = \min\left\{B \in \mathbb{N}_0 : \overline{R}^\star_B - \overline{R}^\star_\infty \le \Delta\right\}. \tag{4}$$

This quantity measures the persistent memory required to approach the prediction quality attainable with unrestricted history.

4 Theory

These questions cannot in general be answered by communication-rate quantities alone, because information transmitted over time and information retained between predictions represent different resource constraints. Persistent memory and communication rate impose different operational constraints. A streaming predictor may transmit only a limited number of bits per update while allowing previously transmitted information to accumulate, whereas a B-bit persistent predictor must repeatedly compress all prediction-relevant information into one of at most 2^B retained states. The distinction becomes particularly visible near the stability boundary: stationary-side rate extrapolation, explosive-side comparators, and the finite numerical rate gap exhibit different behavior around $|a| = 1$ (Fig. 1).

Therefore, rate-based quantities are not generally equivalent to persistent-state requirements. Neither a per-update bit-rate bound nor a cumulative communication budget can be directly converted, through simple normalization, into the required persistent-state capacity. The theory below works directly with finite persistent state and derives memory–risk laws under progressively broader predictive regimes.

4.1 Stationary Fully Observed Predictive Memory

Thus, rate-based quantities do not collapse to a single retained-memory law and cannot in general be interpreted as persistent-state requirements. A rate bound measured in bits per update cannot in

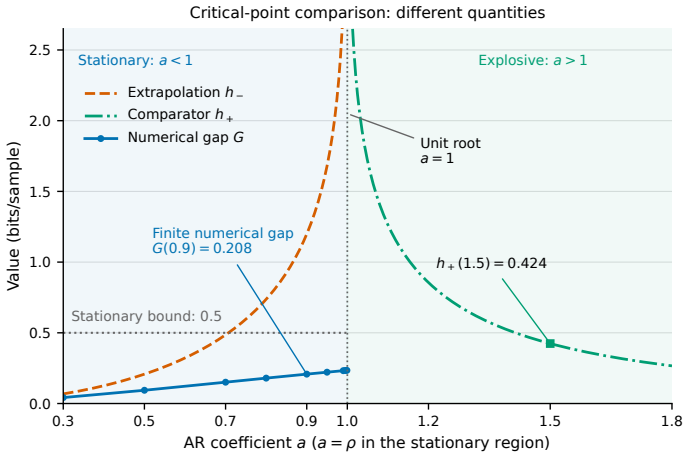

Figure 1: **Rate-based behavior across the stability boundary.** Stationary and explosive-side quantities exhibit distinct behavior near $|a| = 1$.

general be interpreted as a bound on retained memory, and a total communication budget cannot be converted into persistent-state capacity by simple normalization.

Under full observability and stationarity, finite persistent memory reduces to a representation constraint on the current sufficient state. The optimal prediction problem is therefore equivalent to finite-state quantization of that state, and retaining additional history provides no further predictive benefit. Consider the fully observed stationary Gaussian Markov source $X_{t+1} = aX_t + \eta_t$, with $\eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, q)$ and $|a| < 1$. Its stationary variance is $v = q/(1 - a^2)$. For $X \sim \mathcal{N}(0, v)$, let $D_K(v) = \inf_{c_1,\ldots,c_K} \mathbb{E}[\min_j (X - c_j)^2]$ denote the optimal K-point squared-error quantization distortion.

Theorem 1 (Persistent memory collapses to state quantization). For any predictor with at most K persistent states, $R_K^\star = q + a^2 D_K(v)$. Equivalently, $R_B^\star = q + a^2 D_{2^B}(v)$ for a B-bit predictor.

The equality is exact over arbitrary recursive and history-dependent finite-state predictors. No use of earlier observations or previous memory states can improve on directly quantizing the currently observed state X_t.

Indeed, innovation independence gives

$$\mathbb{E}(X_{t+1} - g(S_t))^2 = q + \mathbb{E}(aX_t - g(S_t))^2. \tag{5}$$

For $a \neq 0$, the values $g(S_t)/a$ form at most K reconstruction points for X_t. Conditional on any $X_t = x$, randomization or history-dependent state selection can only form a mixture of squared distances to these points and therefore cannot beat the nearest one. Conversely, storing the label of an optimal K-point quantizer of the current observation attains the bound.

Using the Gaussian information lower bound $D_{2^B}(v) \geq v2^{-2B}$ and the matching one-dimensional quantization rate $D_{2^B}(v) = O(v2^{-2B})$ gives

$$B^\star(d) = \frac{1}{2} \log_2 \frac{a^2 v}{d - q} + O(1), \qquad d \downarrow q. \tag{6}$$

4.2 Dynamic Predictive Memory Complexity

When the current state is observed without noise, the state distribution itself may expand over time. Maintaining a fixed *absolute* prediction accuracy then requires progressively finer representation over an increasingly large dynamic range. Dynamics can create a persistent memory requirement even in the absence of hidden state or partial observability.

Consider again $X_{t+1} = aX_t + \eta_t, \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, q)$, with $X_0 \sim \mathcal{N}(0, v_0)$. Its marginal variance evolves as

$$v_t = a^{2t} v_0 + q \sum_{j=0}^{t-1} a^{2j}. \tag{7}$$

For a fixed target one-step prediction risk $d > q$, let $B_t^\star(d)$ denote the minimum persistent-memory budget attaining risk at most d at time t.

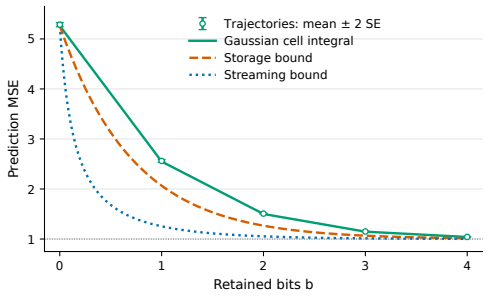

Figure 2: **Finite-state prediction under full observability.**

Theorem 2 (Memory tracks predictive scale). For every fixed t, the optimal B-bit prediction risk is $R_{B,t}^\star = q + a^2 D_{2^B}(v_t)$. Consequently, whenever $v_t \to \infty$,

$$B_t^\star(d) = \frac{1}{2} \log_2 v_t + O(1), \tag{8}$$

where the $O(1)$ term does not depend on t.

Corollary 2 (Three memory-growth regimes). The dynamics induce three qualitatively distinct memory regimes:

$$B_t^\star(d) = \begin{cases} O(1), & |a| < 1, \\ \frac{1}{2} \log_2 t + O(1), & |a| = 1, \\ t \log_2 |a| + O(1), & |a| > 1. \end{cases} \tag{9}$$

The three cases follow respectively from $v_t \to q/(1-a^2)$, $v_t = v_0 + tq$, and $v_t = \Theta(a^{2t})$.

4.3 Inferential Predictive Memory Complexity

Partial observability changes the role of persistent memory qualitatively. The current observation is no longer sufficient for prediction, so a finite-state predictor must preserve information from the past in order to construct a predictive state.

Consider a binary hidden Markov model

$$\Pr(Z_{t+1} \neq Z_t) = \epsilon, \qquad Y_t = Z_t \oplus V_t, \qquad V_t \overset{\text{iid}}{\sim} \text{Bernoulli}(\delta), \tag{10}$$

with $0 < \epsilon, \delta < 1/2$. Let $\alpha = 1 - 2\epsilon$ and $\beta = 1 - 2\delta$. For the stationary long-run analysis, we consider the hidden Markov process in stationarity and define the unrestricted posterior belief from the infinite observation past,

$$p_t = \Pr(Z_t = 1 \mid Y_{-\infty:t}), \qquad \theta_t = \Pr(Y_{t+1} = 1 \mid Y_{-\infty:t}) = \delta + \beta(\epsilon + \alpha p_t). \tag{11}$$

We denote the stationary law of θ_t by μ_θ. Under Brier loss, every finite-state predictor satisfies

$$\mathbb{E}\big[(Y_{t+1} - g(S_t))^2\big] = \mathbb{E}[\theta_t(1-\theta_t)] + \mathbb{E}[(\theta_t - g(S_t))^2]. \tag{12}$$

Thus the entire price of finite persistent memory is the error incurred when representing the Bayes predictive probability θ_t.

Thus finite-memory quality depends not only on how finely belief is represented, but also on how the finite-state dynamics preserve predictive information.

Proposition 1 (Exact one-bit homogeneous predictor). Among all time-homogeneous randomized one-bit state updates with fixed readouts, allowing fresh independent update randomness, the minimum long-run Brier risk is $R^\star_{1,\mathrm{hom}} = \frac{1-\alpha^2\beta^4}{4}$. It is attained by storing the latest observation, $S_t = Y_t$. For arbitrary B, let $\mathcal{E}_B := \overline{R}^\star_B - \overline{R}^\star_\infty$ denote the optimal excess Brier risk.

Theorem 3 (General-B inferential memory bounds). For the stationary hidden Markov process and $B \geq 1$,

$$Q_{2^B}(\mu_\theta) \leq \mathcal{E}_B \leq C_{\epsilon,\delta} \min\left\{\alpha^{2B}, \frac{1}{(1-\alpha)^2(2^B-1)^2}\right\}, \tag{13}$$

where $C_{\epsilon,\delta} = (\alpha\beta H/4)^2$.

Theorem 3 reveals two distinct mechanisms for inferential memory: *history retention* and *belief representation*. The former benefits from rapid Bayesian forgetting and yields an α^{2B} guarantee; the latter exploits finite-state approximation of the predictive belief and yields a 4^{-B}-type guarantee for fixed source parameters. Neither mechanism universally dominates the other. In particular,

$$B^\star_{\mathrm{ex}}(\Delta) \leq \min\left\{\frac{1}{2\log_2(1/\alpha)}, \frac{1}{2}\right\} \log_2 \frac{1}{\Delta} + O(1), \qquad \Delta \downarrow 0. \tag{14}$$

This is an achievability comparison rather than an optimal phase transition: matching the two rates requires additional information about the stationary geometry of the Bayes predictive state.

The resulting memory law is governed by two distinct properties of the predictive process: how rapidly information from the past is forgotten, and how efficiently the resulting predictive state can be represented by a finite set. Predictive memory complexity therefore depends not only on state cardinality, but on the interaction between temporal dynamics and predictive-state geometry.

4.4 Non-Gaussian Predictive Memory Complexity

The Gaussian assumption in Theorem 1 gives a particularly simple quantization law, but the reduction from persistent-memory prediction to finite-state representation does not itself require Gaussianity. Consider the fully observed stationary scalar process $X_{t+1} = aX_t + \eta_t$, where η_t is independent of X_t, has zero mean and variance q, and the stationary law of X_t is P_X with variance v. For $X \sim P_X$, define

$$D_K(P_X) = \inf_{c_1,\ldots,c_K} \mathbb{E}\left[\min_{1\leq j\leq K} (X - c_j)^2\right]. \tag{15}$$

Theorem 4 (Non-Gaussian predictive memory). For any predictor with at most 2^B persistent states,

$$R^\star_B = q + a^2 D_{2^B}(P_X), \qquad R^\star_B - q \quad \geq a^2\beta v\, 2^{-2B}, \tag{16}$$

where the entropy-power factor $\beta = 2^{-2D_{\mathrm{KL}}(P_X\|\mathcal{N}(\mu,v))}$ expresses the KL divergence in bits, consistent with the base-two convention of the appendix. The second inequality holds when P_X is absolutely continuous with finite differential entropy.

For a Gaussian state distribution, $D_{\mathrm{KL}}\big(P_X\|\mathcal{N}(\mu,v)\big) = 0, \beta = 1$, and the lower bound reduces to the Gaussian information bound used in Theorem 1. For non-Gaussian states, β separates variance from distributional geometry: two predictive states with the same variance need not have the same finite-memory difficulty.

Rényi/Sibson information provides a complementary diagnostic of distributional geometry. For Gaussian joint sources with information quantities measured in bits,

$$I_\alpha(X;Y) = \frac{1}{2}\log_2(1 + \alpha\,\mathrm{SNR}), \qquad 2^{2I_\alpha} - 1 = \alpha\,(2^{2I_1} - 1), \tag{17}$$

so the entire α-family is determined by I_1.

4.5 KV-Cache Predictive Memory Threshold

The preceding results characterize predictive memory at the level of finite-state representations. We now translate the same memory–accuracy principle to a stylized KV-cache setting, where a fixed

persistent memory budget must be distributed across both representation dimension and sequence length. Consider the isotropic predictive model

$$X_{t+1} = \rho X_t + \eta_t, \qquad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\eta^2 I_d), \qquad |\rho| < 1, \tag{18}$$

with independent coordinates and $X_t \in \mathbb{R}^d$. The stationary variance of each coordinate is $v = \sigma_\eta^2/(1-\rho^2)$. A cache stores quantized representations $\widehat{X}_t$ of the past L predictive states using a total persistent memory budget of B bits across all dL scalar coordinates. The effective per-scalar memory budget is $R = \frac{B}{dL}$bits per scalar.

The cache-based predictor forms $\widehat{X}_{t+1} = \rho\,\mathbb{E}[X_t \mid \widehat{X}_t]$, which is optimal under squared error for the AR(1) model since $\mathbb{E}[X_{t+1} \mid \widehat{X}_t] = \rho\,\mathbb{E}[X_t \mid \widehat{X}_t]$. The one-step prediction distortion is

$$D_L = \frac{1}{dL} \sum_{t=0}^{L-1} \sum_{i=1}^{d} \mathbb{E}\Big[(X_{t+1,i} - \widehat{X}_{t+1,i})^2\Big], \tag{19}$$

and $D_L^\star$ denotes the infimum of D_L over all cache encoders.

Theorem 5 (KV-cache prediction threshold). Under the isotropic model equation 18, any cache with total memory budget B satisfies

$$D_L^\star \geq \sigma_\eta^2 + \frac{\rho^2 \sigma_\eta^2}{1-\rho^2}\, 2^{-2R}, \qquad R = \frac{B}{dL}. \tag{20}$$

The normalized remaining gap to the innovation floor is

$$\frac{D_L^\star(R) - \sigma_\eta^2}{D_L^\star(0) - \sigma_\eta^2} = 2^{-2R}, \tag{21}$$

and the half-gap threshold is therefore $R_{\text{onset}} = \frac{1}{2}$bits per scalar, $B_{\text{onset}} = \frac{dL}{2}$.

Hence, when $R < R_{\text{onset}}$, the cache retains less than half of the prediction improvement available in the infinite-memory limit. Notably, R_{onset} is *independent of* ρ: regardless of the temporal correlation strength, a per-scalar budget of $\frac{1}{2}$ bit recovers half of the achievable predictive gain. The relevant cache budget is therefore determined by the per-scalar rate, while the *total achievable gain* scales with $\rho^2/(1-\rho^2)$.

To assess whether the proposed stylized cache model covers a plausible range of temporal dependence observed in real Transformer representations, we use GPT-2, a canonical decoder-only Transformer with a standard architecture, transparent attention mechanisms, and layer- and head-wise accessibility, and measure the lag-one temporal correlations of its key representations. GPT-2 (124M) key projections across 144 layer–head pairs exhibit moderate but consistently stable temporal dependence: the mean and median lag-one autocorrelations are $|\rho| = 0.359$ and 0.327, respectively, with an interquartile range of 0.226–0.456. All measured pairs satisfy $|\rho| < 1$, placing them within the stable regime of the stylized cache model. This comparison is intended only to assess the empirical scale of ρ, since real K/V dynamics need not satisfy the coordinate-independent isotropic AR(1) assumption.

5 Implications for Continual World Models

The results above provide quantitative answers to the two questions posed in Section 3. First, Eq. equation 6 characterizes the minimum persistent-state capacity required to attain a prescribed prediction risk. Second, this requirement varies systematically with the predictive process: Eq. equation 8 shows that memory tracks the evolving scale of the predictive state, producing bounded, logarithmic, and linear growth under stable, critical, and unstable dynamics, respectively; Eq. equation 13 quantifies the additional inferential burden induced by partial observability; and Eqs. equation 16–equation 4.4 show how predictive-state geometry modifies finite-memory risk beyond Gaussianity. The relevant resource is the mutable, data-dependent state that can persist across time, rather than communication rate, context length, or model size. Predictive memory complexity therefore characterizes the amount of persistent state required to maintain a prescribed level of prediction accuracy.

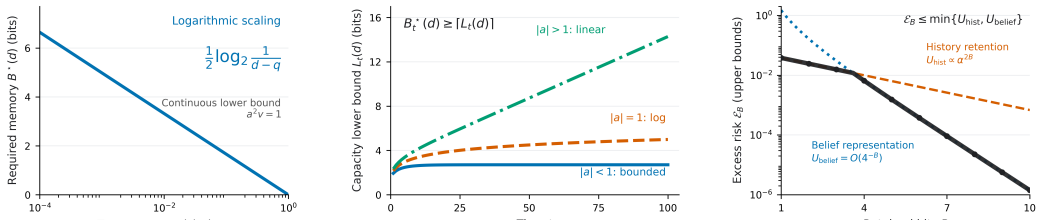

Figure 3: **Three roles of persistent memory.** Memory is representational under full observability and stationarity, dynamic under nonstationary evolution, and inferential under partial observability.

Across the regimes considered above, persistent memory plays qualitatively different roles. Under full observability and stationarity, memory is purely representational: optimal finite-state prediction reduces exactly to quantization of the current predictive state. In the stationary Gaussian example with $a = 0.9$ and $q = 1$, increasing the retained-memory budget from $B = 0$ to $B = 4$ reduces the prediction risk from 5.263 to 1.041, approaching the innovation floor $q = 1$. Thus, in this regime, additional memory primarily increases the resolution with which the current predictive state can be represented.

Under nonstationary dynamics, the required representation capacity becomes time-dependent. Stability determines whether a fixed memory budget remains sufficient: representative cases $|a| = 0.9$, 1, and 1.1 exhibit, respectively, bounded, logarithmic, and approximately linear memory growth. The implication is that memory requirements are determined not only by the desired predictive accuracy, but also by the rate at which the predictive state itself expands over time.

Under partial observability, memory becomes inferential: it must preserve information from history in order to construct the predictive state itself. For the partially observed binary hidden Markov model, finite-memory prediction is governed by two complementary mechanisms. Retaining a recent observation window exploits temporal forgetting, whereas recursively representing the Bayesian belief exploits finite-state approximation of the predictive state. At the smallest budgets, recent-history storage can be more effective, but in the settings considered here the recursive belief representation overtakes it from $B = 3$ onward. Their theoretical bounds show that predictive memory complexity is jointly controlled by filtering contraction and the geometry of the predictive state.

Beyond Gaussian dynamics, the difficulty of representing a predictive state is not determined by variance alone. The non-Gaussian extension shows that distributional geometry enters the memory requirement through an entropy-power correction. Thus two predictive states with the same variance can require different finite-memory resolutions when their distributions have different geometric structure.

6 Conclusion

Predictive memory complexity provides a unified way to characterize the persistent-state resources required for continual prediction. Rather than treating memory as a fixed architectural choice, the results show that its required capacity is determined by the predictive process itself: by the geometry of the predictive state, the scale induced by the dynamics, and the uncertainty introduced when that state must be inferred from history.

This perspective also clarifies what memory means for continual world models. Persistent memory is not simply a larger context window or a larger cache, but the capacity needed to preserve and update the predictive state at the resolution required for future prediction. Consequently, world-model memory should be designed around what predictive information must persist, how accurately it must be represented, and how this requirement changes as the environment evolves. Fixed-capacity memory may therefore be sufficient in some regimes but fundamentally inadequate in others, suggesting that adaptive allocation, compression, and state-update mechanisms should be guided by the structure and evolution of the predictive state rather than by architectural scale alone.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, clarity, and structural refinement of the manuscript. All mathematical arguments, theoretical claims, experimental design, numerical results, and cited sources were reviewed and verified by the authors. The authors take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

The main results are analytical and are reproducible from the complete problem specification and proofs provided in the paper and appendices. Predictive memory complexity, the admissible persistent-state mechanisms, prediction risk, and bit-budget constraints are formally defined. For the Gaussian and nonstationary settings, all process parameters and recursions required to reproduce the theoretical curves and numerical evaluations are specified explicitly. The finite-memory calculations reduce to the stated quantization or rate–distortion quantities, and the procedures used to evaluate these quantities numerically are described together with the relevant parameter ranges. The appendices contain detailed proofs of the main theorems and additional derivations supporting the asymptotic results. The central claims therefore do not rely on proprietary datasets, closed models, or undocumented training procedures.

REFERENCES

Toby Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, Englewood Cliffs, NJ, 1971. ISBN 9780137531035.

Vivek S. Borkar, Sanjoy K. Mitter, and Sekhar Tatikonda. Optimal sequential vector quantization of markov sources. *SIAM Journal on Control and Optimization*, 40(1):135–148, 2001. doi: 10.1137/S0363012999365261.

Ricardo Buitrago and Albert Gu. Understanding and improving length generalization in recurrent models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5856–5874. PMLR, 2025. URL https://proceedings.mlr.press/v267/buitrago25a.html.

Charalambos D. Charalambous, Photios A. Stavrou, and Nasir U. Ahmed. Nonanticipative rate distortion function and relations to filtering theory. *IEEE Transactions on Automatic Control*, 59 (4):937–952, 2014. doi: 10.1109/TAC.2013.2293403.

Kalyan Chatterjee, Konstantin Guryev, and Tai-Wei Hu. Bounded memory in a changing world: Biases in behaviour and belief. *Journal of Economic Theory*, 206:105556, 2022. doi: 10.1016/j.jet.2022.105556.

Xi Chen, Christos Papadimitriou, and Binghui Peng. Memory bounds for continual learning, 2022.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2 edition, 2006. ISBN 9780471241959. doi: 10.1002/047174882X.

Imre Csiszár. Generalized cutoff rates and rényi's information measures. *IEEE Transactions on Information Theory*, 41(1):26–34, 1995. doi: 10.1109/18.370121.

Milan S. Derpich and Jan Østergaard. Improved upper bounds to the causal quadratic rate-distortion function for gaussian stationary sources. *IEEE Transactions on Information Theory*, 58(5):3131–3152, 2012. doi: 10.1109/TIT.2012.2184669.

Meir Feder, Neri Merhav, and Michael Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38(4):1258–1270, 1992. doi: 10.1109/18.144706.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2450–2462, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Martin E. Hellman and Thomas M. Cover. Learning with finite memory. *The Annals of Mathematical Statistics*, 41(3):765–782, 1970. doi: 10.1214/aoms/1177696958.

Martin E. Hellman and Thomas M. Cover. On memory saved by randomization. *The Annals of Mathematical Statistics*, 42(3):1075–1078, 1971. doi: 10.1214/aoms/1177693334.

Michael R. James and Satinder Singh. Learning and discovery of predictive state representations in dynamical systems with reset. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 417–424, 2004. doi: 10.1145/1015330.1015359.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, pp. 1555–1561, 2001. URL https://proceedings.neurips.cc/paper/2001/hash/1e4d36177d71bbb3558e43af9577d70e-Abstract.html.

Daniel Monte and Maher Said. Learning in hidden markov models with bounded memory. Technical Report MSR-TR-2010-79, Microsoft Research, 2010.

Daniel Monte and Maher Said. The value of (bounded) memory in a changing world. Technical Report 47595, Munich Personal RePEc Archive, 2013. URL https://mpra.ub.uni-muenchen.de/47595/.

Girish N. Nair and Robin J. Evans. Stabilizability of stochastic linear systems with finite feedback data rates. *SIAM Journal on Control and Optimization*, 43(2):413–436, 2004. doi: 10.1137/S0363012902402116.

Elvis Nunez, Luca Zancato, Benjamin Bowman, Aditya Golatkar, Wei Xia, and Stefano Soatto. Expansion span: Combining fading memory and retrieval in hybrid state space models. In *Proceedings of the International Conference on Neuro-symbolic Systems*, volume 288 of *Proceedings of Machine Learning Research*, pp. 570–596. PMLR, 2025. URL https://proceedings.mlr.press/v288/nunez25a.html.

Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8733–8744, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Po_Long-Context_State-Space_Video_World_Models_ICCV_2025_paper.html.

Anant Sahai and Sanjoy Mitter. The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link—part i: Scalar systems. *IEEE Transactions on Information Theory*, 52(8):3369–3395, 2006. doi: 10.1109/TIT.2006.878169.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423, 623–656, 1948.

Robin Sibson. Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 14(2):149–160, 1969. doi: 10.1007/BF00537520.

Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, pp. 512–519, 2004.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback–leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

Hans S. Witsenhausen. On the structure of real-time source coders. *Bell System Technical Journal*, 58(6):1437–1451, 1979. doi: 10.1002/j.1538-7305.1979.tb02263.x.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Annan Yu, Michael W. Mahoney, and N. Benjamin Erichson. HOPE for a robust parameterization of long-memory state space models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=RZwtbg3qYD.

Serdar Yüksel. On optimal causal coding of partially observed markov sources in single and multiterminal settings. *IEEE Transactions on Information Theory*, 59(1):424–437, 2013. doi: 10.1109/TIT.2012.2212874.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34661–34710. Curran Associates, Inc., 2023. doi: 10.52202/075280-1506.

Ingvar Ziemann, Nikolai Matni, and George Pappas. State space models, emergence, and ergodicity: How many parameters are needed for stable predictions? In *Proceedings of the 7th Annual Learning for Dynamics and Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pp. 1–11. PMLR, 2025. URL https://proceedings.mlr.press/v283/ziemann25a.html.

A Proofs for Static Predictive Memory Complexity

Resource model and conventions. All information quantities and memory budgets use base-two logarithms; $\ln$ denotes the natural logarithm. A B-bit predictor has at most $K = 2^B$ persistent states, where B is a nonnegative integer. After observing the current input, it updates its state and predicts using only that state. Transition rules, readout tables, and model parameters are fixed read-only specifications, whose implementation costs are separate from the persistent-state budget. No uncounted observation history, real-valued accumulator, or data-dependent random seed is allowed. This is a storage model, not a communication-rate constraint.

We first consider the fully observed stationary Gaussian model

$$X_{t+1} = aX_t + \eta_t, \qquad |a| < 1, \qquad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, q), \qquad X_0 \sim \mathcal{N}(0, v), \quad v = \frac{q}{1-a^2}, \tag{22}$$

where $q > 0$ and the innovations are independent of the initial state and predictor randomness. Updates have the form $S_t = f(S_{t-1}, X_t)$ and readouts have the form $g(S_t)$. Define

$$D_K(v) = \inf_{c_1,\dots,c_K \in \mathbb{R}} \mathbb{E} \min_{1 \le j \le K} (X - c_j)^2, \qquad X \sim \mathcal{N}(0, v). \tag{23}$$

Let $\mathcal{R}_B$ be the infimum long-run mean prediction MSE over admissible predictors, and let $B^\star(d) = \min\{B \in \mathbb{Z}_{\ge 0} : \mathcal{R}_B \le d\}$, with the minimum of an empty set defined as infinity.

Theorem 1 (Exact static reduction)**.** *Under equation 22,*

$$\mathcal{R}_B = q + a^2 D_{2^B}(v). \tag{24}$$

The same expression is the infimum one-step risk at every time, and the infimum cumulative risk over T predictions is $T\mathcal{R}_B$.

Proof. Independence and the zero mean of η_t give the orthogonal decomposition

$$\mathbb{E}(X_{t+1} - g(S_t))^2 = q + \mathbb{E}(aX_t - g(S_t))^2. \tag{25}$$

If $a \neq 0$, put $c_j = g(j)/a$. Pointwise, $(X_t - c_{S_t})^2 \ge \min_j (X_t - c_j)^2$; stationarity and equation 23 therefore imply the lower bound in equation 24, regardless of how S_t uses the past. Conversely,

for any fixed codebook, the update $S_t = \arg\min_j |X_t - c_j|$, with a fixed tie rule, and readout $g(j) = ac_j$ attain the codebook's distortion in equation 25. Taking codebooks arbitrarily close to the infimum proves equality, without assuming that a numerical Lloyd iteration finds a global minimizer. The same codebook works at every time, proving the cumulative assertion. If $a = 0$, the constant predictor zero attains risk q, and the result follows directly. □

[Gaussian finite-cardinality converse] If $X \sim \mathcal{N}(0, v)$ and $\widehat{X}$ takes at most 2^B values, then $\mathbb{E}(X - \widehat{X})^2 \geq v2^{-2B}$.

Proof. Write $D = \mathbb{E}(X - \widehat{X})^2$. Conditional translation, the maximum-entropy property of a Gaussian, and Jensen's inequality give

$$h(X \mid \widehat{X}) = h(X - \widehat{X} \mid \widehat{X}) \leq \frac{1}{2}\log_2(2\pi eD).$$

Thus $I(X; \widehat{X}) \geq \frac{1}{2}\log_2(v/D)$, whereas $I(X; \widehat{X}) \leq H(\widehat{X}) \leq B$. Rearranging proves the claim. Zero or infinite distortion is handled by the same inequality in the extended sense, or by an arbitrarily small regularization. □

[Explicit static bounds] For $B \geq 1$,

$$q + a^2v4^{-B} \leq \mathcal{R}_B \leq q + 5a^2vB4^{-B}. \tag{26}$$

For $B = 0$, $\mathcal{R}_0 = v$. In particular, for $a \neq 0$ and $d > q$,

$$B^\star(d) \geq \left\lceil \left[\frac{1}{2}\log_2 \frac{a^2v}{d-q}\right]_+ \right\rceil, \qquad [x]_+ = \max\{x, 0\}. \tag{27}$$

Proof. The lower bound follows from Lemma A and Theorem 1. For the upper bound, standardize X to $Z = X/\sqrt{v}$. Partition $[-A, A]$ into $K = 2^B$ equal cells, use their midpoints as codewords, and saturate both tails to the extreme codewords. On $|Z| \leq A$, squared error is at most A^2/K^2. On the tails it is at most Z^2. With the standard normal density ϕ and tail Q, integration by parts gives

$$\mathbb{E}(Z - Q_K(Z))^2 \leq A^2K^{-2} + 2\{A\phi(A) + Q(A)\}.$$

Since $Q(A) \leq \phi(A)/A$, choose $A = \sqrt{4B\ln 2}$ to obtain

$$\mathbb{E}(Z - Q_K(Z))^2 \leq \left[4B\ln 2 + \sqrt{\frac{2}{\pi}}\left(\sqrt{4B\ln 2} + \frac{1}{\sqrt{4B\ln 2}}\right)\right] 4^{-B} \leq 5B4^{-B}.$$

The last inequality uses $\sqrt{B} \leq B$ and $1/\sqrt{B} \leq B$ for $B \geq 1$, and $4\ln 2 + \sqrt{2/\pi}(\sqrt{4\ln 2} + 1/\sqrt{4\ln 2}) < 5$. Scaling back proves the upper bound. For one codeword, the optimal reconstruction is zero, so $D_1(v) = v$ and $q + a^2v = v$. Solving the lower risk bound for the integer budget proves equation 27. □

The exact reduction concerns fully observed Markov states. It does not assert that histories are redundant under partial observation. Likewise, equation 26 does not claim matching constants: its explicit uniform-quantizer upper bound contains a factor B.

B Proofs for Dynamic Predictive Memory Complexity

Now allow any $a \in \mathbb{R}$ and $X_0 \sim \mathcal{N}(0, v_0)$ with $v_0 > 0$, retaining independent Gaussian innovations of variance $q > 0$. The source need not be stationary. A time-dependent budget allows $|\mathcal{S}_t| \leq 2^{B_t}$, and a data-independent external clock may select prescribed codebooks and readouts. Such a clock and read-only specifications are not additional data-dependent state. Define $B_t^\star(d)$ as the minimum state capacity at time t allowing one-step MSE at most d, in this full-observation interface.

Theorem 2 (Dynamic converse and growth regimes)**.** *The marginal variance satisfies*

$$v_t = a^{2t}v_0 + q\sum_{j=0}^{t-1} a^{2j} = \begin{cases} a^{2t}v_0 + q(1-a^{2t})/(1-a^2), & a^2 \neq 1, \\ v_0 + tq, & a^2 = 1. \end{cases} \tag{28}$$

Every admissible predictor satisfies

$$\mathbb{E}(X_{t+1} - g_t(S_t))^2 \geq q + a^2 v_t 4^{-B_t}, \tag{29}$$

$$\sum_{t=0}^{T-1} \mathbb{E}(X_{t+1} - g_t(S_t))^2 \geq Tq + a^2 \sum_{t=0}^{T-1} v_t 4^{-B_t}. \tag{30}$$

For $a \neq 0$ and $d > q$, put

$$L_t(d) = \left[\frac{1}{2}\log_2 \frac{a^2 v_t}{d-q}\right]_+. \tag{31}$$

Then $B_t^\star(d) \geq \lceil L_t(d) \rceil$, and

$$L_t(d) = \begin{cases} O(1), & |a| < 1, \\ \frac{1}{2}\log_2 t + O(1), & |a| = 1, \\ t\log_2 |a| + O(1), & |a| > 1. \end{cases} \tag{32}$$

Proof. Induction on $v_{t+1} = a^2 v_t + q$ proves equation 28, and Gaussianity is preserved by the linear recursion. At each fixed time, equation 25 and Lemma A, applied to $g_t(S_t)/a$, prove equation 29. Summing yields equation 30; solving the per-step inequality gives equation 31. For $|a| < 1$, v_t converges to $q/(1-a^2)$. For $|a| = 1$, $\log_2 v_t = \log_2 t + \log_2 q + o(1)$. For $|a| > 1$,

$$v_t = a^{2t}\left(v_0 + \frac{q}{a^2-1}\right) - \frac{q}{a^2-1},$$

so $\log_2 v_t = 2t\log_2 |a| + O(1)$. Substitution proves all regimes. The case $a = 0$ requires no state: the zero predictor has risk q. □

[A constructive dynamic upper bound] For $a \neq 0$, set $V_t = a^2 v_t/(d-q)$ and $s_t = \log_2 \max\{2, V_t\}$. The schedule

$$B_t = \left\lceil \frac{1}{2}s_t + \frac{1}{2}\log_2 s_t + 3 \right\rceil \tag{33}$$

admits an actual finite-label predictor with MSE strictly less than d at every time. Consequently, the proven bounds imply

$$\begin{aligned} B_t^\star(d) &= O(1), && |a| < 1, \\ \tfrac{1}{2}\log_2 t + O(1) \;\leq\; B_t^\star(d) &\leq \tfrac{1}{2}\log_2 t + \tfrac{1}{2}\log_2\log_2 t + O(1), && |a| = 1, \\ t\log_2 |a| + O(1) \;\leq\; B_t^\star(d) &\leq t\log_2 |a| + \tfrac{1}{2}\log_2 t + O(1), && |a| > 1. \end{aligned}$$

The $O(1)$ terms in the two sides need not be the same.

Proof. At time t, quantize $X_t/\sqrt{v_t}$ with the saturated midpoint quantizer from Proposition A, store only its label, and read out $a\sqrt{v_t}$ times its codeword. Its risk is at most $q + 5a^2 v_t B_t 4^{-B_t}$. As $s_t \geq 1$ and $\log_2 s_t \leq s_t$, $B_t \leq s_t + 4 \leq 5s_t$. The lower estimate on the ceiling gives $4^{-B_t} \leq 1/(64\, 2^{s_t} s_t)$, and $V_t \leq 2^{s_t}$. Hence

$$5V_t B_t 4^{-B_t} \leq 25/64 < 1.$$

This proves the risk guarantee. When $V_t \to \infty$, the schedule uses $\frac{1}{2}\log_2 V_t + \frac{1}{2}\log_2\log_2 V_t + O(1)$ bits; equation 28 gives the stated asymptotics. □

For fixed B at $|a| = 1$, the cumulative converse specializes to

$$Tq + 4^{-B}\{Tv_0 + qT(T-1)/2\},$$

and at $|a| > 1$ its per-step excess grows at least as $a^{2t+2}v_0 4^{-B}$. Thus no fixed finite storage budget stabilizes absolute prediction MSE in these regimes. A communication rate threshold does not

contradict this conclusion: it can permit a decoder to retain an uncounted real-valued state. The constructive bounds above assume ideal arithmetic and prescribed codebook scaling; they do not certify fixed-precision implementation over an infinite horizon. The displayed lower-bound curves must not be identified with the exact optimum beyond what the necessary and sufficient bounds establish.

C Inferential Predictive Memory Complexity

Let $Z_t \in \{0,1\}$ be a stationary symmetric Markov chain with flip probability $\varepsilon \in (0,1/2)$, and let $Y_t = Z_t \oplus V_t$, where the V_t are iid Bernoulli(δ), $\delta \in (0,1/2)$, independent of Z. Only Y_t is observed. Write

$$\alpha = 1-2\varepsilon, \qquad \beta = 1-2\delta, \qquad L = \ln\frac{1-\delta}{\delta}, \qquad H = \ln\frac{1-\varepsilon}{\varepsilon} + L.$$

A controller has at most $K = 2^B$ labels, updates from (S_{t-1}, Y_t), and outputs $g(S_t) \in [0,1]$ to predict Y_{t+1} under Brier loss. Fresh source-independent randomization is allowed at each update, but no additional persistent random seed is allowed. Randomized readouts do not improve squared loss: conditional averaging over fresh readout randomness gives a deterministic readout with no larger risk.

Unless stated otherwise, long-run risks mean expected Cesàro averages. For stationary filter calculations, extend the source to all integer times and set $\mathcal{F}_t^Y = \sigma(Y_s : s \le t)$. Finite-start filters give the same limiting bounds by the contraction proved below. We denote the full-history Bayes risk by R_∞ and the infimum long-run excess over B-bit controllers by $\mathcal{E}_B$. The exact one-bit result below is explicitly restricted to time-homogeneous update kernels and fixed readouts; no time-varying optimality is inferred from it.

C.1 Bayes-Risk Decomposition and Filter Contraction

[Risk decomposition] Let $p_t = \mathbb{P}(Z_t = 1 \mid \mathcal{F}_t^Y)$ and

$$\theta_t = \mathbb{P}(Y_{t+1} = 1 \mid \mathcal{F}_t^Y) = \delta + \beta(\varepsilon + \alpha p_t). \tag{34}$$

For every admissible forecast $\widehat{Y}_{t+1}$,

$$\mathbb{E}(Y_{t+1} - \widehat{Y}_{t+1})^2 = \mathbb{E}[\theta_t(1-\theta_t)] + \mathbb{E}(\theta_t - \widehat{Y}_{t+1})^2. \tag{35}$$

For a fixed stationary state process, the optimal readout is $g(s) = \mathbb{P}(Y_{t+1} = 1 \mid S_t = s)$.

Proof. The hidden transition gives prior probability $\varepsilon + \alpha p_t$; the observation channel gives equation 34. Condition on the observation past and independent controller randomness. The conditional expectation of $Y_{t+1} - \theta_t$ is zero, so expansion of the square has zero cross term. The remaining conditional variance is $\theta_t(1-\theta_t)$. Conditional least squares given S_t proves the readout assertion. For reducible controllers, calibration is with respect to the occupation law induced by initialization, not an arbitrarily selected stationary solution. □

[Uniform pathwise contraction] Let $\sigma(\ell) = (1+e^{-\ell})^{-1}$. The log-odds filter is

$$\ell_t = F_{Y_t}(\ell_{t-1}), \qquad F_y(\ell) = \ln\frac{\varepsilon + \alpha\sigma(\ell)}{1-\varepsilon-\alpha\sigma(\ell)} + (2y-1)L. \tag{36}$$

Every update lies in $[-H, H]$, and $|F_y(u) - F_y(v)| \le \alpha|u-v|$. The prediction map $\theta(\ell) = \delta + \beta(\varepsilon + \alpha\sigma(\ell))$ is $\alpha\beta/4$-Lipschitz.

Proof. Bayes' rule multiplies the prior odds by $e^{(2y-1)L}$, proving equation 36. The prior probability lies in $[\varepsilon, 1-\varepsilon]$, which gives the range bound. Put $p = \sigma(\ell)$ and $r = \varepsilon + \alpha p$. Then

$$F_y'(\ell) = \frac{\alpha p(1-p)}{r(1-r)}, \qquad r(1-r) = p(1-p) + \varepsilon(1-\varepsilon)(1-2p)^2.$$

Therefore $0 \le F_y' \le \alpha$. Finally, $\theta'(\ell) = \alpha\beta p(1-p) \le \alpha\beta/4$. The mean value theorem proves both Lipschitz bounds. In particular, two filters driven by the same observations forget a bounded initial log-odds discrepancy at rate α^t. Filters begun arbitrarily far in the past are Cauchy after the first bounded update; their limit is the stationary infinite-past filter used here. □

C.2 Exact Two-Bit Counterexample

[An absorbing four-codeword filter] Take $\varepsilon = 1/50$, $\delta = 3/10$, and nearest-neighbor log-odds quantization on $\{-\ln 27, -\ln 3, \ln 3, \ln 27\}$. For a fixed initial label, its optimal fixed-readout long-run Brier risk is $1/4$. A two-bit register retaining (Y_{t-1}, Y_t) instead attains

$$R_{\mathrm{win},2} = \frac{817896079}{3405437500} < \frac{1}{4}, \qquad \frac{1}{4} - R_{\mathrm{win},2} = \frac{8365824}{851359375}. \tag{37}$$

Proof. For odds r, prediction maps r to $(1 + 49r)/(49 + r)$, and observation zero or one multiplies by $3/7$ or $7/3$. Log-odds midpoints correspond to odds thresholds $1/9, 1, 9$. Exact updated odds and resulting labels are

old label	$r'\ (Y = 0)$	$r'\ (Y = 1)$	new $(Y = 0)$	new $(Y = 1)$
0	$57/2317$	$133/993$	0	1
1	$39/259$	$91/111$	1	1
2	$111/91$	$259/39$	2	2
3	$993/133$	$2317/57$	2	3

No value is a threshold, so the result is independent of tie-breaking. Labels 1 and 2 are absorbing. From label 0 the chain reaches label 1 upon the first observation one; from label 3 it reaches label 2 upon the first observation zero. Either observation has conditional probability at least δ, so the unabsorbed probability after n steps is at most $(1-\delta)^n$. For each fixed initial label, the limiting label is deterministic; its optimal long-run forecast is consequently $1/2$, since the observation marginal is uniform. The finite transient has zero Cesàro contribution. Initializing the real filter at log-odds zero and immediately quantizing after the first observation also gives risk $1/4$: the absorbing label then records only a finite initial observation, whose correlation with future hidden states vanishes geometrically.

For an explicit window calculation, put $A_t = 2Y_t - 1$. The adjacent correlation is $c_1 = \alpha\beta^2$ and the two-step correlation is $c_2 = \alpha^2\beta^2$. Sign symmetry makes odd moments vanish, so the joint law of three signs $U = A_{t-1}, V = A_t, W = A_{t+1}$ is

$$\mathbb{P}(U = u, V = v, W = w) = \tfrac{1}{8}\{1 + c_1 uv + c_1 vw + c_2 uw\}.$$

Hence $\mathbb{E}[W \mid U = u, V = v] = (c_1 v + c_2 u)/(1 + c_1 uv)$. Averaging its square over $\mathbb{P}(U = u, V = v) = (1 + c_1 uv)/4$ yields

$$\mathbb{E}\{\mathbb{E}[W \mid U, V]^2\} = \frac{c_1^2 + c_2^2 - 2c_1^2 c_2}{1 - c_1^2} = c_1^2 + \frac{(c_2 - c_1^2)^2}{1 - c_1^2}.$$

Binary conditional variance therefore gives

$$R_{\mathrm{win},2} = \frac{1 - c_1^2}{4} - \frac{(c_2 - c_1^2)^2}{4(1 - c_1^2)}.$$

Substitution of $\alpha = 24/25$, $\beta = 2/5$ proves equation 37 using rational arithmetic. □

This example compares two controllers with four labels; it does not establish failure of every belief quantizer. In particular, its fixed codebook is different from the endpoint grid in Section C.6.

C.3 Proof of the Exact One-Bit Result

Theorem 3 (Exact one-bit risk)**.** *Among all time-homogeneous kernels on one persistent bit and fixed readouts, allowing fresh independent update randomness, the minimum long-run Brier risk is*

$$R_1^\star = \frac{1 - \alpha^2\beta^4}{4}. \tag{38}$$

Retaining the most recent observation attains this value.

Proof. Write $H_t = 2Z_t - 1$, $A_t = 2Y_t - 1$, and $C_t \in \{-1, 1\}$ for the stored bit. Every binary update kernel has a unique expansion

$$\mathbb{E}[C_t \mid C_{t-1} = \xi, A_t = \zeta] = b + u\xi + v\zeta + w\xi\zeta, \qquad -1 \le b + u\xi + v\zeta + w\xi\zeta \le 1.$$

The finite joint chain (H_t, C_t) has invariant occupation limits; expected Cesàro laws converge even for periodic or reducible chains. Every such invariant law has uniform hidden marginal. Let $c = \mathbb{E}C_t$ and $m = \mathbb{E}H_tC_t$ in that law. Hidden-state transitions and independent emission noise give

$$\mathbb{E}[C_{t-1}A_t] = \alpha\beta m, \quad \mathbb{E}[H_tC_{t-1}] = \alpha m, \quad \mathbb{E}[H_tC_{t-1}A_t] = \beta c.$$

Taking first moments in the update expansion therefore gives

$$(1-u)c - w\alpha\beta m = b, \qquad -w\beta c + (1-u\alpha)m = v\beta. \tag{39}$$

The difference between the two legal corner values at fixed ζ shows $|u + w\zeta| \le 1$, hence $|u| + |w| \le 1$. If $u < 1$, define

$$\Delta = (1-u)(1-u\alpha) - w^2\alpha\beta^2, \quad N = b(1-u\alpha) + w\alpha v\beta^2, \quad J = (1-u)v + wb.$$

Using $w^2 \le (1-u)^2$ gives

$$\Delta \ge (1-u)\{1 - u\alpha - \alpha(1-u)\} = (1-u)(1-\alpha) > 0.$$

Solving equation 39 yields $c = N/\Delta$ and $m = \beta J/\Delta$. For $|c| < 1$, the two possible values of C_t and $\mathbb{E}H_t = 0$ imply

$$\mathbb{E}[H_t \mid C_t] = \frac{m}{1-c^2}(C_t - c), \qquad \mathbb{E}[A_{t+1} \mid C_t] = \frac{\alpha\beta m}{1-c^2}(C_t - c).$$

Thus the optimal calibrated risk for this occupation law is

$$R = \frac{1}{4}\left(1 - \frac{\alpha^2\beta^2 m^2}{1-c^2}\right). \tag{40}$$

The exact polynomial certificate established below gives $\Delta^2 - N^2 - J^2 \ge 0$. It follows that $m^2 \le \beta^2(1-c^2)$, proving the lower bound in equation 38.

The exceptional cases do not require division. If $|c| = 1$, the label is constant in the occupation law, so $m = 0$ and risk is $1/4$. If $u = 1$, legality forces $b = v = w = 0$, so $C_t = C_{t-1}$; any initial hidden-label correlation decays as α^t, again giving long-run risk $1/4$. These arguments cover all invariant occupation limits, including initialization-dependent mixtures of closed classes. Finally $C_t = A_t$ has $c = 0$, $m = \beta$, and readout $g(C_t) = (1 + \alpha\beta^2 C_t)/2$. Equation equation 40 gives equality. □

Exact nonnegativity certificate used in the proof. The argument is computer-assisted only at the following finite algebraic step; it uses exact rational coefficients, not sampled kernels or local optimization. Let $p_0, p_1, p_2, p_3 \in [0,1]$ be the probabilities of output $+1$ at inputs $(-,-), (-,+), (+,-), (+,+)$, respectively. Put $k_i = 2p_i - 1$ and

$$b = \frac{k_0 + k_1 + k_2 + k_3}{4}, \qquad u = \frac{-k_0 - k_1 + k_2 + k_3}{4},$$
$$v = \frac{-k_0 + k_1 - k_2 + k_3}{4}, \qquad w = \frac{k_0 - k_1 - k_2 + k_3}{4}.$$

Write $y = \beta^2$ and $P = \Delta^2 - N^2 - J^2$. Define the nonnegative term

$$Q = \frac{\alpha^2(1-y^2)}{4}\big[(1-p_0)p_3p_1(1-p_2)(p_1+p_2-1)^2 + (1-p_1)p_2p_0(1-p_3)(p_0+p_3-1)^2\big].$$

For $x = (p_0, p_1, p_2, p_3, \alpha, y)$, expand $P - Q = \sum_j a_j x^j$. Its coordinatewise degree vector is $d = (3,3,3,3,2,2)$. Its tensor-product Bernstein expansion is

$$P - Q = \sum_{0 \le i \le d} c_i \prod_{r=1}^{6} \binom{d_r}{i_r} x_r^{i_r}(1-x_r)^{d_r - i_r}, \quad c_i = \sum_{j \le i} a_j \prod_{r=1}^{6} \frac{\binom{i_r}{j_r}}{\binom{d_r}{j_r}}. \tag{41}$$

All 2304 coefficients are nonnegative: 1845 are positive and 459 are zero. Since each basis polynomial and Q are nonnegative on the unit cube, this proves $P \geq 0$. For completeness, the following executable verifier specifies and checks every coefficient with rational arithmetic and independently reconstructs the power-basis polynomial. It requires Python and SymPy; no floating-point tolerance is used.

```
import itertools as it
import sympy as s
p0,p1,p2,p3,A,y = x = s.symbols('p0 p1 p2 p3 A y')
k0,k1,k2,k3 = [2*p-1 for p in x[:4]]
b=(k0+k1+k2+k3)/4; u=(-k0-k1+k2+k3)/4
v=(-k0+k1-k2+k3)/4; w=(k0-k1-k2+k3)/4
D=(1-u)*(1-u*A)-w*w*A*y
N=b*(1-u*A)+w*A*v*y; J=(1-u)*v+w*b
P=D*D-N*N-J*J
Q=A*A*(1-y*y)/4*((1-p0)*p3*p1*(1-p2)*(p1+p2-1)**2
                 +(1-p1)*p2*p0*(1-p3)*(p0+p3-1)**2)
R=s.Poly(s.expand(P-Q), *x)
d=R.degree_list()
assert d == (3,3,3,3,2,2)
indices=list(it.product(*(range(z+1) for z in d)))
c={}
for i in indices:
    c[i]=sum(a*s.prod(s.binomial(ii,jj)/s.binomial(dd,jj)
                      for ii,jj,dd in zip(i,j,d))
             for j,a in R.terms() if all(jj<=ii
                                         for jj,ii in zip(j,i)))
assert len(c)==2304 and all(z>=0 for z in c.values())
assert sum(bool(z>0) for z in c.values())==1845
back=c
for axis,degree in enumerate(d):
    out={}
    for j in indices:
        total=s.S.Zero
        for ii in range(j[axis]+1):
            src=list(j); src[axis]=ii
            total += (back[tuple(src)]*s.binomial(degree,ii)
                      *s.binomial(degree-ii,j[axis]-ii)
                      *(-1)**(j[axis]-ii))
        out[j]=total
    back=out
assert {j:a for j,a in back.items() if a!=0}==R.as_dict()
print('Exact one-bit certificate verified.')
```

The source symmetry alone would only justify a proof for symmetric kernels; the polynomial step is what covers asymmetric kernels as well. No claim of optimality for time-varying controllers or $B > 1$ follows.

C.4 Proof of the General-B Lower Bound

Let μ_θ be the stationary law of equation 34 and define

$$Q_K(\mu_\theta) = \inf_{c_1,\dots,c_K \in [0,1]} \int \min_j (\theta - c_j)^2 \, \mu_\theta(d\theta). \tag{42}$$

Theorem 4 (Finite-image converse)**.** *Every B-bit controller with no extra persistent side information satisfies*

$$R_B \geq R_\infty + Q_{2^B}(\mu_\theta), \qquad \mathcal{E}_B \geq Q_{2^B}(\mu_\theta). \tag{43}$$

The converse also allows deterministic time-dependent readout tables and randomized state updates with fresh independent randomness.

Proof. At each fixed time the readout image $\mathcal{C}_t = \{g_t(s) : s \in \mathcal{S}_t\}$ contains at most 2^B numbers. For every realization, including randomized state updates,

$$(\theta_t - g_t(S_t))^2 \geq \min_{c \in \mathcal{C}_t} (\theta_t - c)^2.$$

Taking expectation and then the infimum over all such codebooks gives $\mathbb{E}(\theta_t - g_t(S_t))^2 \geq Q_{2^B}(\mu_\theta)$. Apply Lemma C.1, average over time, and take an infimum over controllers. Predetermined time variation does not help evade this bound, because the infimum over codebooks is taken separately at each time. For a finite-start formulation, couple the finite and stationary filters to the same observations. Lemma C.1 gives an exponentially vanishing difference in their predictions. On $[0, 1]$, the function $\min_j (\theta - c_j)^2$ is uniformly 2-Lipschitz, independently of the codebook. Therefore the corresponding quantization functionals and Bayes risks converge, and their Cesàro limits give the same lower bound. □

This is an oracle quantization converse, not a realizability theorem: the nearest codeword to θ_t need not be computable from (S_{t-1}, Y_t) with K states. In particular, neither tightness nor a universal 4^{-B} lower scaling follows without further regularity and realizability assumptions on μ_θ and the controller.

C.5 Finite-Window Achievability

Theorem 5 (A B-bit observation window)**.** *For every integer $B \geq 1$, a register storing the last B observations achieves excess risk at most*

$$\mathcal{E}_B \leq \left(\frac{\alpha\beta H}{4}\right)^2 \alpha^{2B}. \tag{44}$$

Proof. The register $S_t = (Y_{t-B+1}, \ldots, Y_t)$ has exactly 2^B possible labels and is updated by dropping its oldest bit and appending Y_t. For each possible register value, precompute

$$\widetilde{\ell}_t = F_{Y_t} \circ \cdots \circ F_{Y_{t-B+1}}(0), \qquad \widetilde{\theta}_t = \theta(\widetilde{\ell}_t).$$

This is a fixed readout table: no real-valued filter is retained across updates. Since the stationary hidden prior at time $t - B$ is uniform, this readout is also the exact conditional probability given that observation window. The infinite-past filter has $|\ell_{t-B}| \leq H$. Applying B contractions on the common observations gives $|\ell_t - \widetilde{\ell}_t| \leq \alpha^B H$. The Lipschitz bound for θ and Lemma C.1 yield

$$\mathbb{E}(\theta_t - \widetilde{\theta}_t)^2 \leq (\alpha\beta/4)^2 \alpha^{2B} H^2.$$

An arbitrary padding convention during the first B steps contributes zero to the long-run average, proving equation 44. □

C.6 Belief-Grid Achievability

Theorem 6 (A recursive endpoint grid)**.** *For $B \geq 1$, let $K = 2^B$ and use the K log-odds codepoints*

$$\mathcal{C} = \left\{-H + \frac{2Hj}{K-1} : j = 0, \ldots, K-1\right\}.$$

Nearest-neighbor recursive quantization on this grid achieves

$$\mathcal{E}_B \leq \left(\frac{\alpha\beta H}{4(1-\alpha)(2^B - 1)}\right)^2. \tag{45}$$

Proof. Initialize at any actual grid codepoint; an extra center at zero is not assumed, because K is even. Let $\widehat{\ell}_t$ denote the codepoint indexed by the stored label and implement

$$\widehat{\ell}_t = Q_{\mathcal{C}}(F_{Y_t}(\widehat{\ell}_{t-1})), \qquad \widehat{\theta}_t = \theta(\widehat{\ell}_t).$$

These equations specify a finite transition table and fixed readout. They do not require storing the exact Bayes filter. Every unquantized update lies in $[-H, H]$, where nearest-neighbor error is at most $h = H/(K-1)$. Coupled to the exact filter, the pathwise error obeys

$$e_t := |\ell_t - \widehat{\ell}_t| \le \alpha e_{t-1} + h, \qquad e_t \le \alpha^t e_0 + \frac{h(1-\alpha^t)}{1-\alpha}.$$

Both initial codepoints and the exact posterior lie in the bounded interval after initialization, so e_0 can be bounded by $2H$. This is a deterministic pathwise bound, and thus controls the second moment directly; no inference from a first-moment bound is needed. By Lemma C.1, the excess at time t is at most

$$(\alpha\beta/4)^2 \left(\alpha^t e_0 + \frac{h(1-\alpha^t)}{1-\alpha}\right)^2.$$

Its limit superior is the right-hand side of equation 45; the same holds for the Cesàro average. Calibration of the readout for the realized occupation law can only improve this guarantee.

Combining Sections C.4–C.6 gives the rigorous inferential sandwich, for $B \ge 1$,

$$Q_{2^B}(\mu_\theta) \le \mathcal{E}_B \le \left(\frac{\alpha\beta H}{4}\right)^2 \min\left\{\alpha^{2B}, \frac{1}{(1-\alpha)^2(2^B-1)^2}\right\}. \tag{46}$$

The two upper bounds correspond to separate feasible controllers, and the minimum selects the better guarantee; it need not identify the better actual risk. All constants depend on the fixed model parameters. The grid constant deteriorates as $\varepsilon \downarrow 0$, so this is not a uniform guarantee over arbitrarily slow hidden-state switching. No equality between the oracle lower bound and either construction, nor any universal optimal memory exponent, is asserted.

D Proofs for Non-Gaussian Predictive Memory Complexity

Consider the fully observed stationary scalar process

$$X_{t+1} = aX_t + \eta_t, \qquad \eta_t \text{ independent of } X_t, \quad \mathbb{E}\eta_t = 0, \quad \mathbb{E}\eta_t^2 = q, \tag{47}$$

whose stationary law P_X has variance v and finite differential entropy (in the absolutely continuous case).

Theorem 7 (Non-Gaussian predictive memory bound)**.** *For any predictor with at most $K = 2^B$ persistent states,*

$$R_B^\star = q + a^2 D_{2^B}(P_X), \tag{48}$$

and, when P_X is absolutely continuous with finite differential entropy $h(X)$,

$$D_{2^B}(P_X) \ge \beta\, v\, 2^{-2B}, \qquad \beta = 2^{-2D_{\mathrm{KL}}(P_X \| \mathcal{N}(\mu, v))}, \tag{49}$$

where D_{KL} is the Kullback–Leibler divergence in bits. The bound is strict when $P_X \ne \mathcal{N}(\mu, v)$.

Proof. The exact reduction equation 48 follows from the same finite-state argument as Theorem 1. Innovation independence gives, for any admissible predictor,

$$\mathbb{E}(X_{t+1} - g(S_t))^2 = q + \mathbb{E}(aX_t - g(S_t))^2. \tag{50}$$

For $a \ne 0$, the values $g(S_t)/a$ form at most K reconstruction points for X_t. Pointwise, $(X_t - g(S_t)/a)^2 \ge \min_j (X_t - c_j)^2$, so taking expectations gives $\mathbb{E}(aX_t - g(S_t))^2 \ge a^2 D_K(P_X)$. Conversely, storing the label of an optimal K-point quantizer attains the bound. The argument uses only independence of η_t from X_t and the variance q, not Gaussianity. If $a = 0$, the zero predictor attains risk q.

For equation 49, let $X \sim P_X$ with variance v and differential entropy $h(X)$ (natural logarithm). Let $\widehat{X}$ be any K-codeword quantizer of X. Since $K = 2^B$, $\widehat{X}$ takes at most 2^B values, so $H(\widehat{X}) \le B$ in bits. By the data-processing inequality, $I(X; \widehat{X}) \le H(\widehat{X}) \le B$.

For the conditional variance $\mathrm{Var}(X \mid \widehat{X})$, the chain

$$\begin{aligned} I(X;\widehat{X}) &= h(X) - h(X \mid \widehat{X}) \\ &\geq h(X) - \frac{1}{2}\log_2\big(2\pi e\,\mathbb{E}[\mathrm{Var}(X \mid \widehat{X})]\big) \end{aligned}$$

uses $h(X \mid \widehat{X}) \leq (1/2)\log_2(2\pi e\,\mathbb{E}[\mathrm{Var}(X \mid \widehat{X})])$, which follows from the maximum-entropy property of the Gaussian applied at each value of $\widehat{X}$ together with Jensen's inequality (concavity of the logarithm). Since $h(X) = (1/2)\log_2(2\pi e\,v)$ for variance v, rearrangement yields

$$I(X;\widehat{X}) \geq \frac{1}{2}\log_2 \frac{v}{\mathbb{E}[\mathrm{Var}(X \mid \widehat{X})]}.$$

Combining with $I(X;\widehat{X}) \leq B$ and exponentiating,

$$\mathbb{E}[\mathrm{Var}(X \mid \widehat{X})] \geq v\,2^{-2B}.$$

Now $\mathrm{Var}(X \mid \widehat{X}) = \mathbb{E}[(X - \mathbb{E}[X \mid \widehat{X}])^2 \mid \widehat{X}]$, so

$$\mathbb{E}[(X - \mathbb{E}[X \mid \widehat{X}])^2] \geq v\,2^{-2B}.$$

For any K-codeword quantizer $\widehat{X}$, $\mathbb{E}[(X - \widehat{X})^2] \geq \mathbb{E}[(X - \mathbb{E}[X \mid \widehat{X}])^2]$, and the infimum over K-codeword quantizers is attained (in the asymptotic sense) when $\widehat{X} = \mathbb{E}[X \mid \widehat{X}]$. Hence $D_{2^B}(P_X) \geq v\,2^{-2B}$.

To extract the entropy-power factor β, recall the characterization Cover & Thomas (2006); Shannon (1948)

$$h(X) = \frac{1}{2}\log_2(2\pi e\,v) - D_{\mathrm{KL}}^{(2)}(P_X \| \mathcal{N}(\mu, v))/\ln 2, \tag{51}$$

i.e., the differential entropy in bits equals $(1/2)\log_2(2\pi e v)$ minus the KL divergence in nats. Rewriting in nats and converting,

$$\begin{aligned} N(X) &= \frac{e^{2h^{(e)}(X)}}{2\pi e} \\ &= v \cdot e^{-2D_{\mathrm{KL}}^{(e)}(P_X \| \mathcal{N}(\mu,v))} \\ &= v \cdot e^{-2(\ln 2)\, D_{\mathrm{KL}}^{(2)}(P_X \| \mathcal{N}(\mu,v))} \\ &= v \cdot 2^{-2D_{\mathrm{KL}}^{(2)}(P_X \| \mathcal{N}(\mu,v))}. \end{aligned}$$

Setting $\beta := N(X)/v$ and using $D_{2^B}(P_X) \geq N(X) \cdot 2^{-2B}$ recovers equation 49. □

Rényi/Sibson identity for Gaussian sources. For a bivariate Gaussian (X, Y) with mutual information I_1 in bits, Rényi α-information in bits reduces to

$$I_\alpha(X;Y) = \frac{1}{2}\log_2(1 + \alpha\,\mathrm{SNR}), \quad \text{where } \mathrm{SNR} = 2^{2I_1} - 1, \tag{52}$$

so that $2^{2I_\alpha} - 1 = \alpha\,(2^{2I_1} - 1)$ and the entire α-family is determined by I_1.

E Proofs for KV-Cache Predictive Memory Threshold

We consider the isotropic predictive model

$$X_{t+1} = \rho X_t + \eta_t, \qquad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\eta^2 I_d), \qquad |\rho| < 1, \tag{53}$$

with independent coordinates and $X_t \in \mathbb{R}^d$. A cache of total budget B bits stores quantized representations $\widehat{X}_t \in \mathbb{R}^d$ for $t = 0, \ldots, L-1$, where each coordinate of $\widehat{X}_t$ uses at most $R = B/(dL)$ bits. Write $v = \sigma_\eta^2/(1-\rho^2)$ for the per-coordinate stationary variance, so $X_{t,i} \sim \mathcal{N}(0, v)$ for all t, i.

Theorem 8 (KV-cache prediction threshold)**.** *For any cache with total budget B and any admissible encoder,*

$$D_L^\star \geq \sigma_\eta^2 + \frac{\rho^2\sigma_\eta^2}{1-\rho^2} \cdot 2^{-2R}, \qquad R = \frac{B}{dL}. \tag{54}$$

The normalized remaining gap

$$\frac{D_L^\star(R) - \sigma_\eta^2}{D_L^\star(0) - \sigma_\eta^2} = 2^{-2R} \tag{55}$$

attains its half-gap at $R_{\text{onset}} = 1/2$ bit per scalar, independent of ρ.

Proof. **Step 1 (one-step decomposition).** For any encoder $\widehat{X}_t = f_t(X_t)$ and the corresponding optimal readout $h(\widehat{X}_t) = \mathbb{E}[X_{t+1} \mid \widehat{X}_t]$, the AR(1) structure gives

$$\begin{aligned}\mathbb{E}\Big[(X_{t+1,i} - h(\widehat{X}_t)_i)^2\Big] &= \mathbb{E}\Big[(\rho(X_{t,i} - \mathbb{E}[X_{t,i} \mid \widehat{X}_t]) + \eta_{t,i})^2\Big] \\ &= \rho^2\,\mathbb{E}[(X_{t,i} - \mathbb{E}[X_{t,i} \mid \widehat{X}_t])^2] + \sigma_\eta^2,\end{aligned} \tag{56}$$

using independence of $\eta_{t,i}$ from $(X_{t,i}, \widehat{X}_t)$ and $\mathbb{E}[X_{t+1,i} \mid \widehat{X}_t] = \rho\,\mathbb{E}[X_{t,i} \mid \widehat{X}_t]$.

Step 2 (coordinate independence). Because the coordinates of X_t are independent and identically distributed, the optimal cache encoder is *coordinate-independent* without loss of optimality: $\widehat{X}_{t,i} = f_i(X_{t,i})$ for some encoder f_i taking at most 2^R values per coordinate. Indeed, cross-coordinate encoding allocates rate budget that, for independent sources, can be redirected to self-encoding without increasing the per-coordinate conditional variance. By the symmetry of i.i.d. coordinates, the rate budget is optimally split as R bits per coordinate.

Under coordinate independence, $\mathbb{E}[X_{t,i} \mid \widehat{X}_t] = \mathbb{E}[X_{t,i} \mid \widehat{X}_{t,i}]$, and the conditional mean $\mathbb{E}[X_{t,i} \mid \widehat{X}_{t,i}]$ takes at most 2^R values. By Lemma A applied coordinatewise to $X_{t,i} \sim \mathcal{N}(0, v)$,

$$\mathbb{E}[(X_{t,i} - \mathbb{E}[X_{t,i} \mid \widehat{X}_{t,i}])^2] \geq v \cdot 2^{-2R}. \tag{57}$$

Step 3 (combine). Substituting equation 57 into equation 56 and averaging over $t \in \{0, \ldots, L-1\}$ and $i \in \{1, \ldots, d\}$,

$$\begin{aligned}D_L^\star &\geq \sigma_\eta^2 + \rho^2 v \cdot 2^{-2R} \\ &= \sigma_\eta^2 + \frac{\rho^2\sigma_\eta^2}{1-\rho^2} \cdot 2^{-2R}.\end{aligned} \tag{58}$$

Step 4 (onset). At $R = 0$, no information is stored. The optimal predictor is then the constant $\mathbb{E}[X_{t+1}] = 0$, giving $D_L^\star(0) = \mathbb{E}[X_{t+1,i}^2] = v = \sigma_\eta^2/(1-\rho^2)$. The maximum achievable improvement over the innovation floor is $D_L^\star(0) - \sigma_\eta^2 = \sigma_\eta^2\rho^2/(1-\rho^2)$, and the fraction of the remaining gap is

$$\frac{D_L^\star(R) - \sigma_\eta^2}{D_L^\star(0) - \sigma_\eta^2} = 2^{-2R}. \tag{59}$$

Setting $2^{-2R} = 1/2$ yields $R_{\text{onset}} = 1/2$ bit per scalar, independent of ρ. □

[Achievability up to a logarithmic factor] Uniform scalar quantization of each $X_{t,i}$ followed by the optimal linear readout attains

$$D_L \leq \sigma_\eta^2 + C\,\frac{\rho^2\sigma_\eta^2}{1-\rho^2}\,R\,2^{-2R}$$

for $R \geq R_0$ and some absolute constants C, R_0. Lloyd–Max quantization matches equation 54 up to a constant.

Proof. For $X_{t,i} \sim \mathcal{N}(0, v)$, uniform scalar quantization with 2^R bins and tail saturation achieves $\mathbb{E}[(X_{t,i} - \widehat{X}_{t,i})^2] \leq C\,v\,R\,2^{-2R}$ for $R \geq R_0$ (see Proposition A). The quantizer output can be

replaced by the per-bin conditional mean at the decoder without additional rate; the resulting distortion is no larger. Substituting into equation 56 and averaging gives the claimed upper bound. The factor of R in the numerator is the standard penalty of uniform quantization; Lloyd–Max quantization removes it at the cost of codebook storage, which is read-only specification and does not enter the per-scalar budget. □

The bound assumes the predictor uses the optimal conditional-mean readout $\widehat{X}_{t+1} = \rho\,\mathbb{E}[X_t \mid \widehat{X}_t]$. For the AR(1) model with independent innovations, this readout is exactly the conditional mean $\mathbb{E}[X_{t+1} \mid \widehat{X}_t]$; any other readout incurs strictly larger MSE under squared error. The bound therefore applies to all cache representations that support this readout. The coordinate-independence restriction in Step 2 is asymptotically tight in R for independent coordinates.